THE PROCEEDINGS OF
THE PHYSICAL SOCIETY

Section B

VOL. 64, PART 6 1 June 1951 No. 378 B

Microscopy by Reconstructed Wave Fronts: II

BY D. GABOR
Department of Electrical Engineering, Imperial College, London

MS. received 11*th September* 1950, *and in amended form* 12*th December* 1950

ABSTRACT. The theory of diffraction microscopy is completed and extended in different directions. In this two-step method of image formation the object is reconstructed by optical means from a diffraction diagram, taken in coherent illumination with light or with electrons. The 'projection method', originally described, and the 'transmission method', recently proposed by Haine and Dyson, are two variants which can be treated by one theory. The process of image formation, the coherence requirements, and the conditions for a good reconstruction are discussed in detail. It is shown that the reconstructed image of extended objects suffers from some spurious detail, but this can be largely suppressed in the 'dark-field' method of reconstruction, in which the illuminating wave is cut out after it has passed through the diffraction diagram.

§ 1. INTRODUCTION

THE principle of diffraction microscopy was described in a previous paper (Gabor 1949, to be referred to as I) with a detailed discussion of one of the schemes for putting it into practice. In this 'projection method' the microscopic object is illuminated by a beam of light or of electrons which issue from a small aperture, either directly or through a lens system without any further lenses between the object and the photographic plate, where it forms a diffraction pattern called a 'hologram'. It was shown that if this photograph, suitably processed, is illuminated by a replica of the original wave, the original object is reconstructed in space, together with a ghost or 'conjugate object', which is in general diffuse and distorted. It has been shown theoretically and experimentally that with objects of a suitable type with sufficiently large clear spaces between the dark parts, the effect of the spurious part of the reconstructed image need not be serious.

Electron microscopy was contemplated from the start as the chief, though not the only application of the method of reconstructed wave fronts. In the course of their experiments, which will be described in a separate paper, Haine and Dyson (1950) became aware of certain considerable practical inconveniences, and proposed a modified optical scheme which they suggested to the author. This is called the 'transmission method', because it operates with an electron microscope of the transmission type, only slightly modified, in the taking of the hologram. While in the projection method the whole electron-optical system is between the small illuminating aperture and the object, in the transmission scheme it is between the object and the photographic plate. This scheme has various advantages, the most obvious being that the object can be inspected and adjusted before taking the hologram by the well-tried methods of ordinary electron microscopy, and must be only defocused to a certain extent for the final exposure. Another equally important advantage is the wide field. In the projection method the illuminating beam had already its final, wide divergence angle, and a correspondingly small cross section in the region where the object could be situated. In the transmission scheme the divergence of the illuminating beam is small, and the field correspondingly large. Moreover, as will be shown later on, these advantages are not obtained at the price of other disadvantages. It was feared originally—and this was the reason why this suggestion was not followed up earlier—that this scheme would require a more complicated optical reconstructing device. But, as will be seen, the apparatus is by no means more complicated than the one required for the projection scheme, moreover it offers the additional advantage that a real image of the illuminating aperture is accessible in it. This makes it possible to cut out the unmodified part of the illuminating wave, the 'background', after it has passed through the hologram, or to shift its phase by a Zernike phase plate, and thus to inspect the reconstructed image by the 'dark-field' or by the phase-contrast method.

The practical modifications which have become necessary naturally suggested reviewing the theory on a somewhat broader basis, and rounding it off by a discussion of operating conditions not dealt with in I.

§ 2. THE PRINCIPLE OF WAVE-FRONT RECONSTRUCTION

Consider a monochromatic, coherent wave with amplitude $U=A\exp(i\psi)$ striking a photographic plate. Assume that the plate is developed by reversal, or printed with an overall 'gamma' Γ[†]. On the 'straight' branch of the Hurter and Driffield curve its amplitude transmission t will be $t=KA^\Gamma=K(UU^*)^{\Gamma/2}$, where K is a constant. Imagine now that U is split into a background or illuminating wave U_0 and a disturbance or secondary wave U_1 due to the object

$$U=U_0+U_1=A_0\exp i\psi_0+A_1\exp i\psi_1=\exp i\psi_0\,[A_0+A_1\exp\{i(\psi_1-\psi_0)\}].$$

On the other hand, if we illuminate the photograph hologram with the background wave alone we obtain immediately behind the emulsion a wave $U_s=tU_0$. The simplest, and also the most advantageous, case is $\Gamma=2$. Assuming this value

$$\begin{aligned}U_s=tU_0=KA^2A_0\exp i\psi_0&=KA_0{}^2\exp i\psi_0\\&\times[A_0+A_1{}^2/A_0+A_1\exp\{i(\psi_1-\psi_0)\}+A_1\exp\{-i(\psi_1-\psi_0)\}].\qquad\ldots\ldots(1)\end{aligned}$$

For simplicity assume that the background is uniform in intensity, i.e. $A_0=$constant, and compare the last two equations. The factors before the brackets are identical, apart from a constant factor. Inside the bracket, we see that $A_0[1+(A_1/A_0)^2]$ has been substituted for the background amplitude A_0. If we assume—and this is a very essential condition—that the background is strong relative to the amplitudes scattered by the object, the change is insignificant. The next terms $A_1\exp i\psi_1$ are identical in the two expressions.

[†] The 'gamma' is, by definition, the slope of the Hurter and Driffield curve, i.e. of the plot of the 'density' (the logarithm of the reciprocal intensity transmission) against the logarithm of the 'exposure' (the product of intensity and time).

This means that the original secondary wave has been reconstructed, with the correct phase, and, neglecting the small term A_1^2/A_0, with the correct relative amplitude. But the last term is new. It represents a wave with the same amplitude as the secondary wave U_1, but with a phase of opposite sign relative to the background.

This is perhaps made clearer by the diagrammatic explanation in Figure 1. If the secondary wave is relatively weak, the resulting wave amplitude is very nearly $\overline{OP}$, which is equal to the background amplitude, plus the in-phase component of the secondary. In the photographic process with $\Gamma=2$ this amplitude, squared, appears as transmission, as shown in the lower figure. The background amplitude is assumed unity, in order to bring the two diagrams to the same scale. The transmitted amplitude is now very nearly the background amplitude, plus *twice* the in-phase component of the secondary wave. The result can be interpreted as shown, i.e. that the reconstructed wave differs from the original only in a 'conjugate amplitude', with equal in-phase and opposite in-quadrature component to the secondary. This might appear an arbitrary interpretation since the in-quadrature component has been suppressed, but it will be shown that it corresponds very closely to the way in which the eye will interpret the secondary amplitude; as a wave issuing from a 'conjugate' or 'ghost' object.

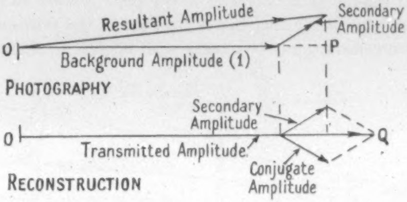


Figure 1. Vector diagram, explaining the principle of reconstruction in the case $\Gamma=2$.

In order to show this we can proceed in a somewhat more general way than in I: we still assume that the illuminating wave issues from a point source, but we drop the assumption, made in I, that the photographic plate is at infinity. The passage to illuminating waves other than spherical can be easily made at a later stage.

Consider a point source S, a point object O_1, and a point P of the photographic plate. P receives a direct ray SP, and an indirect ray SO_1P, with a phase difference between them which is proportional to the difference of the lengths of the broken and of the straight line, plus some phase change which the ray may have suffered in the diffraction at the object. Their interference produces some resulting amplitude at P. But exactly the same amplitude would have been produced with a phase difference of opposite sign between the two rays. We can construct two such rays with opposite phase difference by first of all reversing the direction of propagation, i.e. imagining a spherical wave converging on S instead of diverging from it. Let QO_2 be such a converging wave front. If this is so determined that $\overline{SO_1}+\overline{O_1P}-\overline{SP}=\overline{O_2P}-\overline{PQ}$ the rays from O_1 and from O_2 will have opposite phase difference relative to the direct ray. (If O_1 is not a point, but a small object which produces a certain phase delay, one has to postulate that O_2 produces an equal phase advance.) A simple calculation gives that for small angles α the relation between r_1 and r_2 is

$$1/r_1 + 1/r_2 = 2/R, \qquad \ldots\ldots(2)$$

that is to say the conjugate points O_1 and O_2 are optical conjugates, relative to the spherical wave front through P, considered as a mirror. It may be noted that this is true only in first approximation; for larger values of α, O_2 will depend on α and not only on the radius $R = \overline{SP}$; for constant R it is not a point but an aberration figure, which is not the same as the aberration figure of the spherical mirror. But there are two special cases of particular interest in which the conjugate of a point is a sharp point, without reference to the point P. The first is approached if the object point O_1 is very near to the source S relative to its distance from the plate. This is the *projection case* which was discussed in I; the conjugates are symmetrical with respect to the point source. The second case is realized if the object is very near to the plate relative to its distance from the source. It will be shown that this is the case in the *transmission method*: in this case the conjugates are mirror images with respect to the plate.

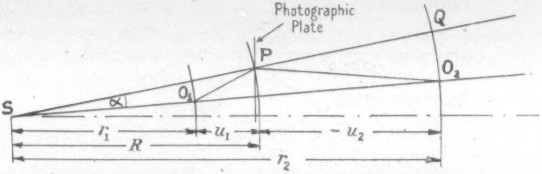


Figure 2. The conjugate object points O_1 and O_2.

One might ask whether some advantage may not be gained by an intermediate arrangement, for instance by placing the object midway between the source and the plate, so that the conjugate is removed to infinity. But nothing is gained by this, because, as may be seen from equation (2), the two conjugate amplitudes at P are always equal. There is no way of getting rid of the conjugate object; it is an unavoidable consequence of the 'phase ignorance' of the photographic plate (cf. Bragg 1950).

Equation (2) can be written, with the notations as in Figure 2,

$$1/u_1 + 1/u_2 = 2/R.$$

This is an optically invariant relation, that is to say if in the reconstruction one images the optical space of the source and the hologram in such a way that the spherical wave from the source to the plate has some other curvature, the two conjugates remain mirror images relative to the transformed spherical wave front.

In equation (1) it was assumed that the wave used in the reconstruction has the same phase differences between different points of the photographic plate as the one used in the taking. We can now drop this assumption, and show that, at least in this approximation, the substitution of one illuminating wave for another has the same effect as an optical transformation. Assume, as illustrated in Figure 3, that the plate was originally illuminated by S′ from a distance R', while in the reconstruction this is changed to R''. Imagine now that we place before the hologram a thin negative lens, with a power $-1/F=1/R'-1/R''$ and a positive lens with equal and opposite sign behind it. Thus the total effect is *nil*, but as the negative lens has restored the original curvature of the wave front, it is seen immediately that in the optical space between the two lenses the conjugates have their original positions, O_1' and O_2'. Hence the substitution of the wave front has the same effect as the positive lens behind the hologram, and we obtain for instance for the first point the transformation equation

$$1/u_1'-1/u_1''=1/F=1/R'-1/R''.$$

In particular, if the new illuminating wave is plane, the hologram acts like a lens with power $1/R'$ in transforming the original position of the object. This, with some other interesting optical properties of holograms has been recently pointed out by Rogers (1950).

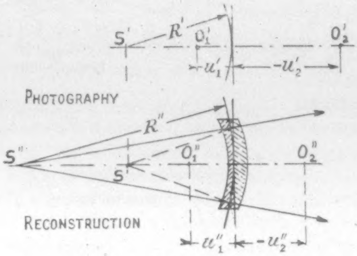


Figure 3. Substituting a different illuminating wave in the reconstruction has the same effect as an optical transformation.

If the illuminating wave is not spherical the conjugate object is no longer sharp, but marred by *twice* the number of aberrations corresponding to the deviations of the wave front from a sphere. This is illustrated in Figure 4 (Plate*). For the theory we must refer to I.

§ 3. THEORY OF IMAGE FORMATION IN THE TRANSMISSION METHOD

Figure 5 illustrates the optics used in the 'projection' and 'transmission' methods. In both cases three electron lenses are required to produce the necessary high demagnification of the pinhole in the first scheme, and the high magnification of the object in the second. It can be seen immediately that the transmission method is a sort of virtual projection method, if the last optical space is considered, which contains the photographic plate, in the H-plane. But it will be shown that it is simpler to consider instead the first optical space, which contains the pinhole and the object. This leads in a much more straightforward way to a simple understanding of the reconstructing process, and it will also show the essential identity of the two methods, which we have emphasized in advance by choosing the same symbol z_0 for the source–object distance in the first scheme, and for the defocusing distance in the second.

The notations are explained in Figure 6. H_0 is the gaussian conjugate of the H-plane, i.e. of the photographic plate; it is a small distance z_0 off the object plane, and will be reckoned positive as illustrated, i.e. with objective 'overfocused'.

---

* For Plates see end of issue.

The calculation consists of two essential steps. In the first we calculate the amplitudes in the H_0-plane, which we call the 'virtual hologram'. In the second we take into account the aberrations of the optical system and calculate the actual hologram. The first may be called $U^0(x_0, y_0)$, the second $U(X, Y)$.

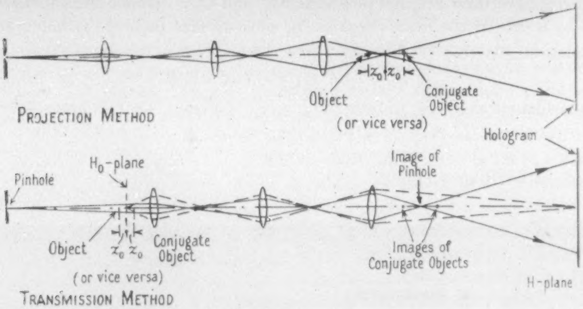

z_0 z_0
Object
Conjugate Object
(or vice versa)
PROJECTION METHOD
Hologram
H_0-plane
Pinhole
Image of Pinhole
Object
z_0 z_0
Conjugate Object
(or vice versa)
TRANSMISSION METHOD
Images of Conjugate Objects
H-plane

Figure 5. Electron-optical system in the two methods of producing the hologram.

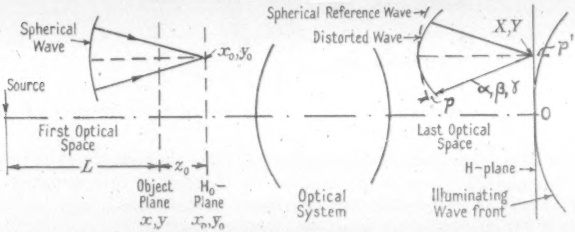

Spherical Wave
Source
x_0, y_0
First Optical Space
L
z_0
Object Plane x, y
H_0-Plane x_0, y_0
Optical System
Spherical Reference Wave
Distorted Wave
X, Y
P'
α, β, γ
P
O
Last Optical Space
H-plane
Illuminating Wave front

Figure 6. Notations for the theory of the transmission method. The z-axis is the optic axis. α, β, γ are in turn the angles of the normals of the plane waves with the x, y, z axes. In the geometrical approximation they become rays.

To calculate the first we use Kirchhoff's principle in the simplified form valid for small off-axis angles

$$U^0(x_0, y_0) = \frac{1}{z_0}\iint U_0{}^0(x, y)\,t(x, y)\exp(ikr_1)\,dx\,dy. \qquad \ldots\ldots(3)$$

Here $U_0{}^0(x, y)$ is the illuminating amplitude in the object plane, $t(x, y)$ is the amplitude transmission coefficient of the object, in general a complex quantity, $k = 2\pi/\lambda$ is the propagation constant, and r_1 is the distance of the point of reference x_0, y_0, z_0 in the H_0-plane from a point $x, y, 0$ in the object plane. We expand

$$r_1 = z_0 + \tfrac{1}{2}[(x-x_0)^2 + (y-y_0)^2]/z_0 - \tfrac{1}{8}[(x-x_0)^2 + (y-y_0)^2]^2/z_0{}^3 + \ldots.$$

It can be verified *a posteriori* that the defocusing distances z_0 suitable for practical use are large enough to justify the application of Kirchhoff's formula in the simplified form (3), moreover the third term in the expansion of r_1 can be dropped. We are *a fortiori* entitled to drop the corresponding term in the phase of the illuminating wave $U_0{}^0$, which is, apart from a constant factor,

$$U_0{}^0 = \exp \tfrac{1}{2}ik(x^2 + y^2)/L.$$

We thus obtain for the amplitude in the virtual hologram

$$U^0(x_0, y_0) = \exp\left(\frac{ik}{2z_0} r_0^2\right) \iint t(x, y) \exp\left[\tfrac{1}{2} ik\left(\frac{1}{z_0} + \frac{1}{L}\right) r^2 - i\frac{k}{z_0}(xx_0 + yy_0)\right] dx\, dy, \qquad \ldots\ldots(4)$$

where we have used the abbreviations $x^2 + y^2 = r^2$ and $x_0^2 + y_0^2 = r_0^2$.

This virtual hologram is imaged on the photographic plate through the optical system, with certain ray limitations, and with certain geometrical aberrations. These can be described, as is well known, by specifying the phase distortion of a wave which is spherical in the first optical space and converges on a point x_0, y_0 of the H_0-plane.

To simplify the formulae we assume the magnification to be unity. This is no restriction, as ultimately it will be convenient to refer all optical data back to the object space. A plane wavelet, which is a component of the distorted wave in the last optical space (corresponding to the spherical wave in the first space which converges on x_0, y_0), will arrive at a point X, Y with a phase advance

$$(x_0 - X)\cos\alpha + (y_0 - Y)\cos\beta + p(\alpha, \beta, x_0, y_0) + p'(X, Y).$$

Here we have split the phase distortion into two components, illustrated in Figure 6; p is the distance, measured in radial direction, between the distorted wave and a spherical wave of reference which coincides with it on the radius drawn parallel to the axis. This is a function of the angular variables α, β and of the point x_0, y_0. The second term p' is the advance of the wave as a whole. This is a function of X, Y only, and represents the distortion of a wave which was plane and normal to the axis in the first space. But as it depends on X, Y only, it adds only a phase factor $\exp(ikp')$ to the amplitude, which has no effect on the photographic density, and hence on the whole process. Therefore we can drop p' from the start in all the following formulae.

We neglect image distortion, field curvature and third-order astigmatism, but take into account first-order astigmatism, spherical aberration and coma. It is well known that astigmatism on the axis, arising from ellipticity or misalignment of the electrodes, is a very important error in electron lenses, and in view of this it would not be justified to assume the 'spherical' aberration to be rotationally symmetrical. We only assume for simplicity that its principal axes coincide with the astigmatic axes, and write

$$p = \tfrac{1}{2} A_s(\cos^2\alpha - \cos^2\beta) + \tfrac{1}{4}(C_x \cos^4\alpha + 2C_{xy}\cos^2\alpha\cos^2\beta + C_y\cos^4\beta)$$
$$+ B(x_0\cos\alpha + y_0\cos\beta)\sin^2\gamma.$$

A_s is the axial separation of the two astigmatic foci, C_x, C_y and C_{xy}, which have the dimension of a length, are the constants of the aperture error; in the case of rotational symmetry $C_x = C_y = C_{xy} = C_s$ is the constant of the spherical aberration. B is the coma coefficient.

The effect of the aperture in the optical system may be represented by a transmission factor $\exp\{-(\gamma/\gamma_m)^2\}$. Such a 'gaussian' cut-off is analytically very convenient, as has been shown in I. It will be also convenient, as in I, to replace the angular variables α, β, γ by 'Fourier variables' $\xi = (\cos\alpha)/\lambda$, $\eta = (\cos\beta)/\lambda$, $\rho = (\xi^2 + \eta^2)^{1/2} = (\sin\gamma)/\lambda$.

With these conventions and notations, applying again Kirchhoff's formula in the simplified form, we obtain for the amplitude in the H-plane

$$U(X, Y) = \iint U^0(x_0, y_0)\, dx_0\, dy_0 \\ \times \iint \exp\{-(\rho/\rho_m)^2 - 2\pi i[(x_0 - X)\xi + (y_0 - Y)\eta + p]\}\, d\xi\, d\eta.$$

It is easy to check by Fourier's formula that, if there is no distortion, i.e. $p = 0$, and if there is no ray-limitation, this transformation restores U^0, that is, in this case $U(X, Y) = U^0(X, Y)$.

We now substitute U^0 from equation (4) and obtain the amplitude in the hologram in the form of a sixfold integral

$$U(X, Y) = \iiiint\!\!\iint t(x, y) \exp\left\{\frac{\pi i}{\lambda}\left[\left(\frac{1}{z_0} + \frac{1}{L}\right)(x^2 + y^2) \\ + \frac{1}{z_0}[x_0{}^2 + y_0{}^2 - 2(xx_0 + yy_0)]\right]\right\} \exp\{-2\pi i[(x_0 - X)\xi + (y_0 - Y)\eta \\ + \tfrac{1}{2}A_s\lambda(\xi^2 - \eta^2) + \tfrac{1}{4}\lambda^3(C_x\xi^4 + 2C_{xy}\xi^2\eta^2 + C_y\eta^4) \\ + B\lambda^2(x_0\xi + y_0\eta)\rho^2]\} \exp[-(\rho/\rho_m)^2]\, dx\, dy\, dx_0\, dy_0\, d\xi\, d\eta. \quad \ldots\ldots(5)$$

But this can be immediately reduced to a fourfold integral, because x_0 and y_0 occur in the exponentials only in the first and second order, and this part of the integral can be readily calculated by the formula

$$\iint \exp \frac{i\pi}{\lambda z_0}(x_0{}^2 + y_0{}^2) \exp\left\{-2\pi i\left[x_0\left(\frac{x}{\lambda z_0} + \xi(1 + B\lambda^2\rho^2)\right) \\ + y_0\left(\frac{y}{\lambda z_0} + \eta(1 + B\lambda^2\rho^2)\right)\right]\right\} dx_0\, dy_0 \\ = i\lambda z_0 \exp\left[-\pi i\lambda z_0\left\{\left(\frac{x}{\lambda z_0} + \xi(1 + B\lambda^2\rho^2)\right)^2 + \left(\frac{y}{\lambda z_0} + \eta(1 + B\lambda^2\rho^2)\right)^2\right\}\right]. \quad \ldots\ldots(6)$$

Hence

$$U(X, Y) = \iiiint t(x, y) \exp\left[\pi i\left(\frac{r^2}{\lambda L} - \lambda z_0\rho^2\right)\right] \exp\left[-\left(\frac{\rho}{\rho_m}\right)^2\right] \\ \times \exp\{-2\pi i[(x - X)\xi + (y - Y)\eta]\} \exp\{-2\pi i[\tfrac{1}{2}A_s\lambda(\xi^2 - \eta^2) \\ + \lambda^3[\tfrac{1}{4}(C_x\xi^4 + 2C_{xy}\xi^2\eta^2 + C_y\eta^4) - Bz\rho^4] + B\lambda^2(x\xi + y\eta)\rho^2]\}\, dx\, dy\, d\xi\, d\eta. \quad \ldots\ldots(7)$$

This is the simplest form in which the exact * solution can be put. It gives the amplitude in an image, focused or defocused, in any microscope, ordinary or electronic, with a gaussian aperture, up to errors of the third order, for stigmatic, coherent illumination of the object from a distance L.

The interpretation is simple if we take the factors one by one. Comparing with (5) the coma now appears transferred to the object plane, i.e. x, y appear in it instead of x_0, y_0. The spherical aberration has suffered a change, which in

---

* It is exact, strictly speaking, only in so far as Kirchhoff's formula can be considered as exact. But this approximation is valid in electron optics with an accuracy almost unknown in light optics, as the wavelength of fast electrons is small compared with the dimensions of any material object except atomic nuclei.

the case of rotational symmetry is equivalent to a decrease of C_s by $4Bz_0$. This will be seen later to be absolutely negligible. Thus the phase distortion due to aberrations, in the last factor of equation (7) is appreciably the same as if the system were focused on the object. The only appreciable effect of the defocusing appears in the factor $\exp(-\pi i\lambda z_0\rho^2)$ which may be called the 'defocusing factor'.

Instead of operating with the amplitude U, it is advantageous, as in I, to operate with a 'shadow object' which, placed in the hologram plane and illuminated with the background U_0, would produce the amplitude U immediately behind it. This 'shadow object' has the *complex* transmission coefficient $\tau(X, Y) = U(X, Y)/U_0(X, Y)$, which differs from the hologram in the imaginary part only. The background amplitude U_0 can be calculated by the method of stationary phase, and is

$$U_0(X, Y) = \frac{L}{L+z_0} \exp\left\{2\pi i\left[\frac{X^2-Y^2}{2\lambda(L+z_0)} - \tfrac{1}{2}A_s\frac{X^2-Y^2}{\lambda(L+z_0)^2} - \frac{\tfrac{1}{4}(C_xX^4+2C_{xy}X^2Y^2+C_yY^4)-B(L+z_0)R^4}{\lambda(L+z_0)^4}\right]\right\}, \quad \ldots\ldots(8)$$

where we have written $X^2+Y^2=R^2$. The terms of the exponential after the first are found to be absolutely negligible, and even the first is small. In practice one can consider the illumination in the transmission method as plane, but we will retain the first term, for some later conclusions.

A considerable further simplification of equation (7) is possible only in the case when the defocusing is large compared with the astigmatic separation A_s and with the length of the caustic in the spherical aberration figure, i.e. if $z_0 \gg A_s$ and $z_0 \gg C_s\gamma_m^2$. In this case a somewhat lengthy calculation gives the formula for the transmission of the 'shadow object'

$$\tau(X, Y) = \frac{1}{i\lambda z_0}\exp\left(\frac{\pi i}{\lambda z_0}R^2\right)\iint t(x, y)\exp\left(\frac{\pi i}{\lambda z}r^2 - \frac{W^2}{(z_0\gamma_m)^2}\right) \times \exp\left[-\frac{2\pi i}{\lambda z_0}(Xx+Yy)\right]\exp\left\{-2\pi i\left[\tfrac{1}{2}A_s\frac{(X-x)^2-(Y-y)^2}{\lambda z_0^2} + \tfrac{1}{4}C_s\frac{W^4}{\lambda z_0^4} + B\frac{[X(X-x)+Y(Y-y)]W^2}{\lambda z_0^3}\right]\right\}dx\,dy, \quad \ldots\ldots(9)$$

where we have introduced the abbreviation $W^2=(X-x)^2+(Y-y)^2$. The passage from (7) to (9) can be expressed by the simple rule: Replace every pencil of wave normals converging in X, Y by a single *ray* from x_0, y_0 to X, Y. It represents therefore the geometrical approximation to the solution. A lengthy investigation, which may be omitted, shows that this approximation is in fact valid almost up to the point when z_0 approaches the tip of the caustic. A further approximation, in which the 'ray' is drawn not simply through x_0, y_0, but is determined by the condition of stationary phase, is valid practically without restriction in diffraction microscopy where the geometrical errors in the caustic region strongly outweigh the diffraction spread of the wave. This however requires the solution of a cubic equation for every ray, and the explicit solution is too complicated to be discussed with advantage.

It may be noted that if the geometrical errors are neglected, equation (9) becomes identical with the equation (14) of I which gives the 'shadow transform' of an object, illuminated by a point source from a distance z_0, with the plate at infinity, apart from the ray-limiting factor $\exp[-W^2/(z_0\gamma_m)^2]$. But it is of course just the ray limitation which constitutes the chief difference between the projection method and the transmission method. In the projection method a point of the photographic plate receives radiation from *all* the object, if there is sufficiently fine detail present, while in the transmission method it is, broadly speaking, only a disc of radius $z_0\gamma_m$ in the object which contributes to the amplitude at one point of the hologram. Hence, in spite of the formal similarity, which we have emphasized by using the same symbol z_0 in its two different meanings in both papers, there is a very essential difference in the actual holograms.

For a general discussion we can as well start from the exact formula (7) as from the approximation (9). For simplicity put $L=\infty$, i.e. assume parallel illumination. In this case the background U_0, given by equation (8), is reduced to unity, and $U(X,Y)$ gives the transmission of the shadow object as well as the amplitude at the hologram. The transmission of the photograph, assuming correct processing, as explained in § 1, is proportional to the real part of τ, hence, in this case, of U. But it can be seen on inspection of equation (7), with $L=\infty$, that taking the real part is equivalent to *two* objects, with transmissions $t(x,y)$ and $t^*(x,y)$ at $\pm z_0$ from the plane H_0, and reversing the sign of the aberrations, A_s, C_x, etc. for the second. This is illustrated in Figure 7, which shows the position of the spherical aberration caustics of a point object and its conjugate. After the reconstruction one of the objects will be sharp—it does not matter which of the two—and the other will be affected by twice the aberrations. This is a point to which we will return later.

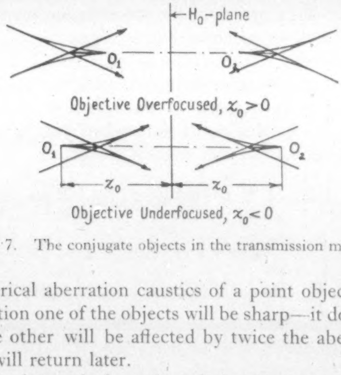

H_0-plane
O_1
O_2
Objective Overfocused, $z_0>0$
O_1
O_2
z_0
z_0
Objective Underfocused, $z_0<0$

Figure 7. The conjugate objects in the transmission method.

It may be instructive to look at the relation between object and hologram from a different point of view. Note that in equation (7), which gives the amplitude in the hologram, the coordinates x, y of the object plane occur only in an integral

$$\iint t(x,y)\exp\frac{\pi i r^2}{\lambda L}\exp[-2\pi i(x\xi+y\eta)(1+B\lambda^2\rho^2)]\,dx\,dy. \qquad ......(10)$$

For simplicity put again $L=\infty$, parallel illumination, and replace ξ, η by the new variables

$$\xi'=\xi(1+B\lambda^2\rho^2); \qquad \eta'=\eta(1+B\lambda^2\rho^2). \qquad ......(11)$$

The expression (10) now appears in the form

$$\iint t(x,y)\exp[-2\pi i(x\xi'+y\eta')]\,dx\,dy,$$

which is the Fourier transform of $t(x,y)$ in the new variables ξ', η', in the standard notation of Campbell and Foster (1931). Writing Ft for the Fourier transform, we have thus for (10) the expression $Ft[\xi(1+B\lambda^2\rho^2),\ \eta(1+B\lambda^2\rho^2)]$. One can say that the coma has produced a certain distortion of the transform in Fourier space. Substituting this into (7), where for simplicity we write the aperture error in the rotationally symmetrical form, we obtain

$$\begin{aligned} U(X,Y) = \iint & Ft[\xi(1+B\lambda^2\rho^2),\ \eta(1+B\lambda^2\rho^2)] \exp\{-2\pi i[\tfrac{1}{2}\lambda z_0\rho^2 \\ & + \tfrac{1}{2}A_s\lambda(\xi^2-\eta^2) + \tfrac{1}{4}C_s\lambda^3\rho^4]\} \exp[-(\rho/\rho_m)^2] \exp[2\pi i(X\xi+Y\eta)]\, d\xi\, d\eta. \end{aligned}$$

This again is a Fourier integral in the standard form (a 'left-transform' if (10) is called a 'right-transform'). Inverting it, we obtain the Fourier transform of U, to be written FU,

$$FU(\xi,\eta) = Ft[\xi(1+B\lambda^2\rho^2),\ \eta(1+B\lambda^2\rho^2)] \exp[-(\rho/\rho_m)^2] \exp[-2\pi i(\tfrac{1}{2}\lambda z_0\rho^2+P)] \qquad \ldots\ldots(12)$$

where P is the phase advance due to astigmatism and spherical aberration only; the coma is contained in the factor Ft. The result means that the Fourier transform of the amplitude at the hologram is obtained by simple multiplication of the (coma-distorted) Fourier transform of the object. In other words, a simply periodic component of the object transmission of density $\exp 2\pi i(x\xi'+y\eta')$ will be transformed in the H-plane into another simply periodic component $\exp[2\pi i(X\xi+Y\eta)] \exp[-(\rho/\rho_m)^2 - 2\pi i(\tfrac{1}{2}\lambda z_0\rho^2+P)]$, where the connection between ξ, η and ξ', η' is given by (11). Provided that there are no aberrations other than those here considered, this is true in every plane, whether the image is focused or not. Defocusing merely shifts the phases of the different components by the factor $\exp(-\pi i\lambda z_0\rho^2) \simeq \exp[-\pi i\lambda z_0\rho'^2(1-2B\lambda^2\rho'^2)]$, where in the second expression we have made use of the fact that in all practical cases $B\lambda^2\rho^2 = B\gamma^2 \ll 1$.

In spite of its formal simplicity this Fourier interpretation of an optical transformation is in general only of limited usefulness—though Duffieux (1950) has given many interesting examples to the contrary—because the intensity has to be calculated at every point by summing the amplitudes of the Fourier components and taking their absolute square. But this is different in diffraction microscopy if it is properly carried out, that is, if the amplitude of the uniform background is large compared with all other amplitudes. A uniform background $U_0^0(x,y)=1$ is by (12) transformed into $U_0(X,Y)=1$, also uniform, and the leading term in the intensity is simply the sum of the real parts of the Fourier components. Thus we obtain a new view of diffraction microscopy, as reconstruction from the ***real part*** of the periodic components of the amplitude in the hologram. In a special case this has been already indicated in I.

§ 4. THE RECONSTRUCTION

We have already seen in § 1 that it is not necessary to illuminate the hologram in the reconstruction with an exact replica of the original wave. This is of great importance in the transmission method. As we have seen in the last section, the hologram can be considered as obtained with the background, such as it was, in the first optical space; only the geometrical and ray-limitation errors of the optical system need be taken into consideration, but not the phase distortion p' of the background. As the illumination in the first optical space is substantially parallel, we can illuminate the hologram in the reconstruction by a plane wave. But we must of course correct the first-order astigmatism, spherical aberration and the coma. It may be noted that because of the mirror symmetry of the conjugates in the first optical space we can use a correcting system with errors of the same or opposite sign to that of the original errors, depending on which conjugate one wants to correct.

So far we have assumed the magnification as unity. We now give it the value M, which can be obtained electronically, in the microscope, or by optical magnification of the hologram. Let us give, as in I, one prime to the parameters used in the taking, e.g. λ' for the de Broglie wavelength, two primes to those in the reconstruction, e.g. λ'' for the wavelength of light. We can now obtain all parameters of the reconstruction by inspection of equation (7), or of the simpler equation (9), by postulating that all phases, measured in periods or in fringes, must be the same in the reconstruction as in the taking. We thus obtain for instance from the first factor before the integral sign in (9), the condition $R^2/\lambda' z_0' = M^2 R^2/\lambda'' z_0''$ or $z_0'' = M^2 z_0' \lambda'/\lambda''$.

It is convenient to introduce the 'excess magnification' m by $m = M\lambda'/\lambda''$. This indicates by how much the actual magnification exceeds the ratio of the light and electron wavelengths. In practical applications this will be probably of the order unity, as λ''/λ' is about 100,000 and this order of magnification is required for high-resolution photographs. With this notation we obtain, by inspection of (7) or (9)*,

$$z_0'' = m^2 z_0' \lambda''/\lambda'\,; \quad A_s'' = m^2 A_s' \lambda''/\lambda'\,; \quad C_s'' = m^4 C_s' \lambda''/\lambda'\,; \quad B'' = m^2 B'. \quad \ldots\ldots(13)$$

The spherical aberration constant of electron microscope objectives C_s' is of the order of a few centimetres, hence with m unity we obtain a value of C_s'' of the order of several kilometres. But it is not of course necessary to carry out the correction in the optical space of the hologram. If one introduces, behind the hologram, a demagnifying optical system, in which m is for instance 1/10, the spherical aberration constant in this space comes down to reasonable dimensions.

Thus the reconstruction apparatus will have to include a demagnifier, at least if M exceeds 10,000. Moreover it is advantageous to make this demagnification variable in certain limits. As the spherical aberration required for correction varies with the fourth power of the magnification, a variation of m in the ratio 1 : 2 is sufficient to cover a range of C_s' from 1 to 16.

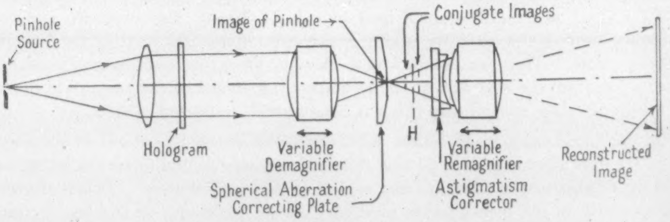


Figure 8. Optical reconstructing device.

Figure 8 is a schematic illustration of a reconstructing device on these lines. The hologram is illuminated by a parallel beam. This is brought to a point focus by means of a variable demagnifying system. The accessibility of this point focus is an important advantage for several reasons. One is that, as Dyson has pointed out in connection with reconstruction devices for projection holograms, this focus is the proper place for a fourth-order correcting plate; elsewhere the plate would introduce strong coma, which would have to be corrected. The other is that in this scheme it is possible to cover up the point focus, preferably with a dark spot graded according to a probability law, and thus use the dark-field method, which will be discussed at the end of this paper. Coma correction can be effected by shifting the fourth-order plate slightly in axial direction.

---

* These formulae have been obtained also in I, but note that z_0 had of course a different meaning in the projection method.

In the optical space after the demagnifier the conjugate images appear approximately at equal distances from the image H of the hologram. The astigmatism can be corrected by two cylindrical lenses of opposite sign, each of which can be rotated. These are followed by a remagnifier, also variable, which produces the final image on a photographic plate.

Instead of using the hologram, processed with $\Gamma=2$ in transmission, one could also process it with $\Gamma=1$ and back it by a mirror, that is use it in reflection. This allows the first lens to be used both for collimation and for demagnification, but the advantages of this scheme are somewhat doubtful, especially in view of the loss of light introduced by a beam-splitting device.

Summing up, the reconstructing device in the transmission method is by no means more complicated than in the projection scheme, and it has the added advantage of an accessible real image of the light source, which could have been realized in the projection scheme only by the introduction of almost prohibitive complications.

§ 5. COHERENCE CRITERIA

In the theory we have so far assumed a monochromatic point source which emits absolutely coherent radiation. The effects of replacing this by a real, small source have been discussed on general lines, and in the special case of the projection method in I. It will be necessary to apply the theory to the transmission method, and also to discuss the question, most important practically, of intensity and exposure.

One can distinguish three coherence criteria, which may be characterized as 'transverse', 'longitudinal' and 'chromatic'. The first concerns the admissible diameter of the pinhole source, the second the constancy of position of the H_0-plane, and the third the spectral width of the radiation.

Transverse Coherence

We can obtain the criterion for the admissible source diameter from equation (7), which gives for the amplitude in the hologram plane due to an infinitesimal area of the object

$$u(X, Y)\,dx\,dy = t(x, y)\exp\left(\frac{\pi i}{\lambda L}r^2\right)dx\,dy$$
$$\times \iint \exp(-\pi i\lambda z_0\rho^2)\exp\{-2\pi i[(x-X)\xi+(y-Y)\eta+p]\}\,d\xi\,d\eta. \quad \ldots\ldots(14)$$

The background amplitude is, on the other hand, by (8)

$$U_0(X, Y) = \exp[2\pi i(R^2/2\lambda L + \;.\;.\;.)],$$

where we have quoted only the relevant part, the rest being entirely negligible. For satisfactory coherence we require that the phase of u does not change by more than half a fringe, that is by π, if the illuminating spot moves in a circle of diameter d_s. Instead of moving the source we can as well move the object plane and the H_0-plane by equal amounts, that is we add equal amounts to x and X, or to y and Y. It is seen that the expression under the integral sign in (14) depends only on $x-X$ and on $y-Y$, and hence remains unchanged. (The coma makes a negligible difference, as may be seen if equation (9) is used instead of (7).) Thus the phase difference between u and U_0 which arises from the moving of the source is $(\pi/\lambda L)(r^2-R^2)=(\pi/\lambda L)[(x-X)(x+X)+(y-Y)(y+Y)]$. The worst case is if x, y and X, Y are in line with the axis 0, 0 and the movement of the spot, of value d_s is also in the same line. In this case we obtain the criterion

$$(2d_s/\lambda L)\,|r-R| \leqslant 1. \qquad \ldots\ldots(15)$$

Now $|r-R|$ must be interpreted as the radius of the circle in the H_0-plane in which the information about the point-object is contained, down to the detail d_A, where $d_A=\lambda/2\gamma_m$ is the Abbe resolution limit. If the defocusing distance z_0 is large compared with A_s and the length of the spherical aberration caustic $C_s\gamma_m^2$, the information can be considered as contained in a cone of angle γ_m, so that $|r-R|=z_0\gamma_m$. This, substituted into (15), gives the coherence criterion

$$d_s \leqslant \frac{L}{z_0}\frac{\lambda}{2\gamma_m} = \frac{L}{z_0}d_A \quad (z_0 \ll A_s,\, C_s\gamma_m^2). \qquad \ldots\ldots(16)$$

If however z_0 is small we must take into consideration that the information is contained in a region which is essentially that of the geometrical aberration figure, because under the conditions of diffraction microscopy the geometrical errors can be considered as large compared with the diffraction spread on the beam. Considering for simplicity only spherical aberration, the radius of this region is $C_s\gamma_m^3$ in the gaussian plane, and $\frac{1}{4}C_s\gamma_m^3$ in the plane of minimum confusion. Substituting this into (15) we obtain a new criterion, valid for small z_0-s, which it is not necessary to write down explicitly. It is sufficient to remark that from the point of view of transverse coherence the effective defocusing distance z_0 in equation (16) can never be less than one-quarter of the length of the caustic, i.e. $z_0 \geqslant \frac{1}{4}C_s\gamma_m^2$. This may also be stated in the form that the pinhole of diameter d_s, seen from the object, must never subtend an angle larger than the wavelength divided by the diameter of minimum confusion.

Longitudinal Coherence

One of the most important practical difficulties in diffraction microscopy, or, for that matter, of any method of improving the resolving power of electron microscopes, is the required high constancy of focus. It may be remembered that electron microscopy operates with focal lengths of the same order as optical microscopy, a few millimetres, while achieving resolutions about 100–200 times better. Moreover electron lenses are not as stable as glass lenses, but are subject to fluctuations, and they are not achromatic. Hence magnetic electron microscopy has become possible only by very careful stabilization of the lens currents, to about one part in 20,000. This is not so critical in 'unipotential' electrostatic microscopes, where the focal length remains fixed. Even this was possible only through the relatively very large focal depth of electron objectives, due to the small aperture angles. But any further progress sharpens the requirements regarding stability, as the focal depth decreases with the square of the aperture angle.

The limits for the variation of z_0, that is of the gaussian plane which is focused, can be immediately seen from equation (14). This contains z_0 only in the factor $\exp(-\pi i\lambda z_0\rho^2)$ under the integral sign. The effect of the variation of z_0 by Δz is greatest when ρ is greatest. If z_0 itself is sufficiently large, the maximum effective value of ρ is $\rho_m = \gamma_m/\lambda$. Postulating that Δz shall cause a phase shift by less than half a fringe we obtain the *sufficient* criterion

$$\Delta z \leqslant \pm \tfrac{1}{2}\lambda/\gamma_m{}^2 = d_A/\gamma_m, \qquad \ldots\ldots(17)$$

which means simply that the variations must remain inside the focal depth. This criterion is sufficient *and* necessary for large $z_0 - s$. There is no need to investigate the somewhat less stringent necessary conditions for small $z_0 - s$, where the object merges into the caustic, because, as will be seen later, such a setting has considerable practical disadvantages.

The practical consequences of the criterion (17) will be discussed by Haine in a separate publication, with particular reference to the magnetic electron microscope. But it may be noted that even with absolute stabilization of the lenses, or with unipotential electrostatic systems, a limit will be reached, at about 1 to 2 A. resolution, where the energy spread of the electrons will bar further progress, unless achromatic lenses are used. The possibilities of achromatic electron lenses have been discussed by the author in a separate paper (Gabor 1951).

Chromatic Coherence

The criterion for chromatic coherence can also be obtained from equation (14), in which the most important λ-dependent factor is again, if z_0 is not too small, $\exp(-\pi i\lambda z_0\rho^2)$. Applying the same criterion as before, we obtain for the largest admissible relative change of wavelength $|\Delta\lambda/\lambda| \leqslant \lambda/z_0\gamma_m{}^2 = 2d_A/z_0\gamma_m$. Even with the best resolutions contemplated, and the largest $z_0 - s$ which may be compatible with exposures of reasonable length, this gives values well inside the limits to which present-day electron microscope supplies are stabilized. There is therefore no need to take this criterion into consideration, except if achromatized unipotential electrostatic lenses are used, so that the criterion (17), which otherwise overshadows it, is eliminated.

The Coherent Current

Consider the scheme of illumination, illustrated in Figure 9. The information on a very small object is contained, in the H_0-plane, in a disc of radius $z_0\gamma_m$. In order to operate the method correctly, this area must receive a coherent primary wave. This is the condition formulated in equation (16). (In I it is shown that this condition actually corresponds to 72·3% coherence.) It will now be shown that the electron current through this disc, the 'coherent current', is entirely limited by the conditions of emission.

We read from Figure 9, combined with equation (16), the relations $\gamma_s = z_0\gamma_m/L = (d_A/d_s)\gamma_m$ and combining this with Abbe's relation we obtain

$$\gamma_s d_s = \gamma_m d_A = \tfrac{1}{2}\lambda. \qquad \ldots\ldots(18)$$

This means that in coherent beams the Smith–Lagrange invariant has a definite value, equal to half the wavelength. We now apply this to electrons. Calling the velocity V, and the maximum transversal velocity component V_t we have $\gamma_s = V_t/V$. On the other hand, by de Broglie's relation, $\lambda = \boldsymbol{h}/\boldsymbol{m}V$. Substituting this into (18) we obtain $d_s = \boldsymbol{h}/2\boldsymbol{m}V_t$. This is quite generally true for any cross section, if we put in, at the right-hand side, the maximum tranverse velocity in that cross section. The area is $S_c = \frac{1}{4}\pi d_s^2 = \frac{1}{16}\pi(\boldsymbol{h}/\boldsymbol{m}V_t)^2$, where c stands for coherent.

Well-known methods are available in electron optics for the limitation of V_t, hence it might appear that we can increase, for instance, the coherent area of the cathode beyond any limit. But it will now be shown that if the maximum transverse velocity is reduced, the current approaches a definite limiting value. Assume at the cathode surface a Maxwellian distribution with a charge density

$$C \exp(-\boldsymbol{m}v^2/2\boldsymbol{k}T)\, dv_x dv_y dv_z = 2\pi C \exp\{-\boldsymbol{m}(v_n{}^2 + v_t{}^2)/2\boldsymbol{k}T\} v_t dv_t dv_n.$$

If V_t is the maximum transverse velocity admitted into the beam, the current through the area S_c is

$$I_c = \frac{\pi}{16}\left(\frac{\boldsymbol{h}}{\boldsymbol{m}V_t}\right)^2 2\pi C \int_0^\infty \exp(-\boldsymbol{m}v_n{}^2/2\boldsymbol{k}T) v_n dv_n \int_0^{V_t} \exp(-\boldsymbol{m}v_t{}^2/2\boldsymbol{k}T) v_t dv_t. \qquad \ldots\ldots(19)$$

We express this as the product of an 'effective coherently emitting area' S_{eff} with the total emission density of the cathode, that is, we put

$$I_c = S_{\text{eff}} 2\pi C \int_0^\infty \exp(-\boldsymbol{m}v_n{}^2/2\boldsymbol{k}T) v_n dv_n \int_0^\infty \exp(-\boldsymbol{m}v_t{}^2/2\boldsymbol{k}T) v_t dv_t. \qquad \ldots\ldots(20)$$

From (19) and (20)

$$S_{\text{eff}} = \frac{\pi}{16}\left(\frac{\boldsymbol{h}}{\boldsymbol{m}V_t}\right)^2 \frac{\int_0^{V_t} \exp(-\boldsymbol{m}v_t{}^2/2\boldsymbol{k}T) v_t dv_t}{\int_0^\infty \exp(-\boldsymbol{m}v_t{}^2/2\boldsymbol{k}T v_t dv_t} = \frac{\pi}{16}\left(\frac{\boldsymbol{h}}{\boldsymbol{m}V_t}\right)^2 \{1 - \exp(-\boldsymbol{m}V_t{}^2/2\boldsymbol{k}T\}$$
$$\leqslant \pi\boldsymbol{h}^2/32\,\boldsymbol{mk}T,$$

the limit being reached when the cut-off is at very small velocities, so that $\boldsymbol{m}V_t{}^2/2\boldsymbol{k}T \ll 1$. This means that the maximum coherently emitting area of a thermionic cathode depends on nothing but its temperature.

This area is very small, and so are the coherent currents I_c which can be obtained from it, as shown by the Table for tungsten cathodes.

Temperature T (° K.)	2400	2500	2600	2700	2800	2900	3000
Emission density (amp/cm^2)	0·116	0·298	0·717	1·63	3·54	7·31	14·1
Max. area S_{eff} (10^{-14} cm^2)	1·42	1·36	1·31	1·26	1·21	1·17	1·14
Coherent current I_c (10^{-14} amp.)	0·165	0·405	0·94	2·06	4·30	8·55	16·1

It may be noted that the coherent current is quite independent of the resolution limit. But its limitation makes itself very strongly felt at high resolutions, for two reasons. In the H_0-plane this current is spread over an area $\pi(z_0\gamma_m)^2 = (\pi/4)(\lambda z_0/d_A)^2$, hence the current density in the object space decreases with the square of the resolution, if the defocusing distance z_0 is held constant. For reasons explained in the next section it may be even necessary to increase z_0 with increasing resolving power, hence conditions may be even worse. Second, the magnification must be increased proportionally to the resolving power, if d_A is to be kept in some given proportion to the grain size of the photographic plate, which gives a further quadratic factor. Thus the exposure times will have to be increased, other things being equal, with at least the fourth power of the resolution. We shall not discuss this in detail, as this will be done in a separate paper by Mr. M. E. Haine, but need only remark that with tungsten cathodes and the best photographic plates available the required exposure times become of the order of an hour if a resolution of better than about 2 A. is aimed at.

It is therefore very desirable to find emitters with higher emission density than tungsten cathodes. Autoelectronic emission is an obvious suggestion. Benjamin and Jenkins (1940) found that current densities at least 1,000 times greater than the thermionic emission of tungsten could be maintained for long times. This would reduce the exposure times to seconds instead of hours. Considerable development work will, however, be needed to adapt these very sensitive point cathodes to demountable devices and, as their velocity spread is probably of the order of some volts, they will probably be useful only in conjunction with achromatic lenses, which form the subject of a special investigation.

§ 5. CONDITIONS FOR THE TAKING OF HOLOGRAMS

Assuming that the electron source, the supply voltage and the magnification are suitably fixed, the only important remaining parameter is the defocusing distance z_0. There are several considerations which determine this, of which we mention three.

The first consideration is intensity. We have already seen that, for example, the current density at the object is inversely proportional to the square of z_0, though it cannot be increased beyond any limit, because with decreasing z_0 the 'effective z_0' never falls below one-quarter of the length of the caustic. Thus it would be desirable to immerse the object into the aberration figure. Such sharp focusing is difficult and precarious, and it conflicts also with the two other considerations.

It is the essence of the method of diffraction microscopy that as much information as possible must be put into the clear spaces, where the background is strong, that is it must appear in the form of diffraction fringes surrounding the object. (These are always called Fresnel fringes by electron microscopists though light opticians do not seem to approve of the term.) A typical object suitable for diffraction microscopy, with relatively large clear spaces of average width D, is shown in Figure 10. The clear space is best utilized if the fringes from both sides just cover it, that is if $z_0\gamma_m \simeq \frac{1}{2}D$, or, using Abbe's relation $z_0 \simeq Dd_\Lambda/\lambda$. For example, if $D = 1000$ A. and $d_\Lambda = 5$ A., z_0 is 10 microns, and for $d_\Lambda = 1$ A. z_0 is only 2 microns. It can be already seen that at high resolutions this condition becomes unimportant.

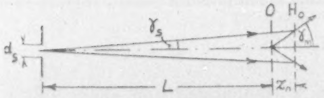


Figure. 9 Illustrating the concept of the coherent current.

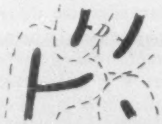


Figure 10. Typical object suitable for diffraction microscopy.

A third condition for z_0 is obtained from considering the reconstruction, in which the conjugate object must disturb the reconstructed object as little as possible. If one image is made sharp, the other appears with aberrations doubled, as shown in Figure 11 for the case of spherical aberration. Two cases must be distinguished, $z_0 > 0$ and $z_0 < 0$. Theory (cf. Picht, 1931, p. 163) and the optical experiments illustrated in Figure 4 show clearly that the second is the more favourable case. In front of the pointed end of the caustic there extends a region of appreciable length in which the intensity has a rather sharp maximum on the axis. In the case $z < 0$, on the other hand, the point O_1 is in the hollow region of the aberration figure of its conjugate O_2, where the intensity has a flat minimum at the axis. The reconstruction in Figure 4 was obtained with the object in this region, experiments with the opposite position failed because of very large density differences in the photograph, which went far beyond the straight region of the Hurter and Driffield curve.

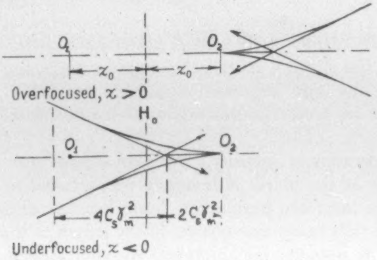


Figure 11. The conjugate objects in the reconstruction, after correcting the errors in one of them.

Within a region which has twice the length of the axial caustic this region is surrounded by a bright annular region, containing a fine system of interference fringes. These vanish at a certain point, and one can see from Figure 11 that the condition for this is

$$-z_0 > 3C_s\gamma_m^2 = \tfrac{3}{4}C_s(\lambda/d_A)^2. \qquad \ldots\ldots(21)$$

As an example let $C_s = 1$ cm. (magnetic microscope), $\lambda = 0{\cdot}05$ A. (corresponding to 60 kev.) and $d_A = 5$ A. Then $z_0 > 0{\cdot}75\,\mu$, much less than the criterion stated above. But with $d_A = 1$ A. the new criterion gives $z_0 > 19\,\mu$, which is much more. It is seen that the condition (21) becomes stringent only at high resolutions. It appears advisable always to take the larger of the two figures.

It may be asked whether reconstruction is not possible from a sharply focused photograph. *One* sharp image can, of course, be always obtained, but this is of little use if it is overlaid by the diffuse image of the conjugate. Even the reconstruction of high-contrast objects is doubtful, because in the region of the caustic high-contrast objects produce a complicated and sharp system of interference fringes, which can be expected to make identification very difficult.

§ 6. THE DARK-FIELD METHOD OF RECONSTRUCTION

As already explained in I, a hologram has the advantage over an ordinary photograph, taken in incoherent illumination, that it can be treated as a physical object rather than as a photograph, and can be inspected in bright-field, dark-field or with the phase-contrast method.

It has been also shown that small objects, with dimensions smaller than the characteristic length $(\lambda z_0)^{1/2}$ give reconstructions in which the disturbance by the conjugate or ghost is not serious, and may even be negligible. But in electron microscopy this condition is not always easily satisfied. As a sufficiently transparent support is not available, the object must be stable in itself during the long exposures, which is not always the case with thin filamentary objects as illustrated in Figure 10.

One of the least favourable cases, which occurs frequently, is the reconstruction of the more or less straight and sharp edge of an extended object. For simplicity we will discuss the case of an absorbing half-plane only, bounded by a straight edge. The reconstruction gives one sharp image, but behind this, at a distance $2z_0$, appears the 'ghost plane', with an extended Fresnel-fringe system, which may be so contrasty that it masks even the images of smaller objects which project from the edge, and which, in themselves, would be very suitable objects for diffraction microscopy. It will now be shown that this very unfavourable case can be greatly improved by the dark-field method. As already explained, in this method the direct or illuminating wave is removed after it has passed the hologram, by placing a small, preferably 'apodized' dark spot at a real image of the pinhole source. The 'apodization', i.e. grading, removes the diffraction fringes which might arise at a sharply limited spot.

For simplicity we disregard the geometrical aberrations in the following calculation. The expression (7) for the amplitude in the hologram now simplifies to

$$U(X,Y)=\iiiint t(x,y)\exp\left[-(\rho_{\mathrm{m}}^{-2}+\pi i\lambda z_0)\rho^2\right] \\ \times\exp\{-2\pi i[(x-X)\xi+(y-Y)\eta]\}\,dx\,dy\,d\xi\,d\eta,$$

where we have separated the Fourier factor under the integrand. The transmission in the hologram, at least in those regions which are neither too much underexposed, nor overexposed, is proportional to $U+U^*$, where U^* is the complex conjugate of U.

The amplitude in the reconstruction, in the plane of one of the images, is obtained by the same formula, by adding the 'refocusing factor' $\exp(\pi i\lambda z_0\rho^2)$ under the integral sign. Applying this to $U+U^*$ we obtain an amplitude

$$U_{\mathrm{rec}}=\iiiint t(x,y)\exp\left[-(\rho^2/\rho_{\mathrm{m}}^2)\right]\exp\left[-2\pi i(x-X)\xi+(y-Y)\eta\right]dx\,dy\,d\xi\,d\eta \\ +\iiiint t(x,y)\exp\left[-(\rho_{\mathrm{m}}^{-2}-2\pi i\lambda z_0)\rho^2\right]\exp\{2\pi i[(x-X)\xi+(y-Y)\eta]\}\,dx\,dy\,d\xi\,d\eta. \qquad \ldots\ldots(22)$$

The first line represents the reconstructed object, the second the ghost.

We now assume the object to be a half-plane, so that $t=0$ from $x=-\infty$ to 0, and $t=1$ for positive values of x. It will however be simpler to make $t=\pm\frac{1}{2}$ for $x\gtrless 0$, and add the uniform level $\frac{1}{2}$ afterwards. The integrals in (22) now become well-known Fourier integrals (cf. Campbell and Foster 1931) and we obtain for the first expression in (22)

$$\tfrac{1}{2}+\tfrac{1}{2}\operatorname{erf}(\pi\rho_{\mathrm{m}}X)=\tfrac{1}{2}+\pi^{-1/2}\int_0^{\pi\varrho_{\mathrm{m}}X}\exp-z^2dz \qquad \ldots\ldots(23)$$

and for the second, assuming $\lambda z_0\rho_{\mathrm{m}}^2\gg 1$,

$$\tfrac{1}{2}+\tfrac{1}{2}\operatorname{erf}\left[\frac{\pi\rho_{\mathrm{m}}X}{(1-2\pi i\lambda z_0\rho_{\mathrm{m}}^2)^{1/2}}\right]=\tfrac{1}{2}+\tfrac{1}{2}(1+i)\left[\mathrm{C}\left(\frac{\pi X^2}{2\lambda z_0}\right)-i\mathrm{S}\left(\frac{\pi X^2}{2\lambda z_0}\right)\right] \qquad \ldots\ldots(24)$$

where erf is the error function, and C, S are the Fresnel integrals.

This result is however, as was said before, valid only in the range of moderate exposures, where the background can be considered as strong. If the half-plane is completely absorbing, there is no background on the shadow side, and the amplitude is obtained by taking the absolute square of the 'edge wave', which is the difference between (24) and its value without diffraction, that is for $\lambda=0$. Thus the intensity at the shadow side is $\frac{1}{4}[(C+S-1)^2+(C-S)^2]$ while on the bright side the intensity is obtained, without appreciable error, by adding the real part of (24) to the background, and squaring, giving $[\frac{3}{4}+\frac{1}{4}(C+S)]^2$. The intensities thus obtained are plotted in Figure 12. It has been assumed for simplicity that the resolution limit is very small compared with the characteristic length $(\lambda z_0)^{1/2}$, so that the correctly reconstructed part of the image, corresponding to (23) is a step-function.

It can be seen from Figure 12, and this has been also found in numerous experiments, that the reconstruction is by no means satisfactory. The fringes at the bright side are too prominent. One can suppress them by contrasty photography, but this suppresses also useful detail.

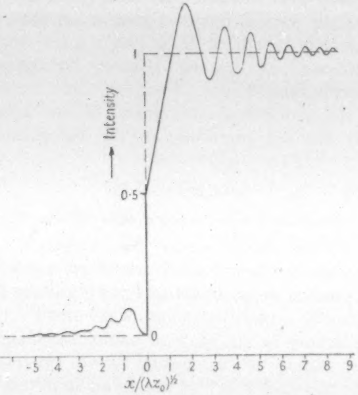


Figure 12. Reconstruction of a straight edge in the bright-field method.

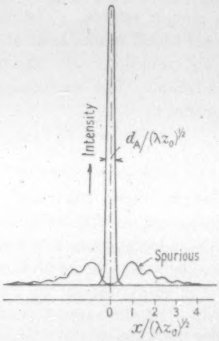


Figure 13. Reconstruction of a straight edge in the dark-field method.

But if the background is suppressed, the picture changes completely. There is no amplification of the fringes, the intensity, as shown in Figure 13, consists of the superposition of the two edge waves, one focused, the other unfocused, both with equal light sums, that is equal areas under the curves. But if the resolution limit is small compared with $(\lambda z_0)^{1/2}$, the intensity in the sharp image can be enormously larger than the spurious image. Thus the correct outline of the object becomes visible, with very high contrast, and under suitable conditions the image may not be noticeably different from a dark-field image in which the ghost is entirely absent. Thus by combining bright-field, dark-field, and perhaps phase-contrast observations, it may be possible to observe much true detail even in otherwise rather unsuitable objects. This is well borne out by provisional observations in optical experiments.

ACKNOWLEDGMENT

A part of this paper was written while the author was a member of the Research Laboratory of The British Thomson-Houston Company Ltd., Rugby, and the photographs in Figure 4 were taken in collaboration with Mr. I. Williams. The author wishes to thank Mr. L. J. Davies, Director of Research, the British Thomson-Houston Company, for permission to publish these photographs.

REFERENCES

Benjamin, M., and Jenkins, R. O., 1940, *Proc. Roy. Soc.* A, **176**, 262.

Bragg, W. L., 1950, *Nature, Lond.*, **166**, 399.

Campbell, G. A., and Foster, R. M., 1931, *Fourier Integrals for Practical Applications, Bell Telephone System, Monograph* B 584.

Duffieux, P. M., 1950, *Réunion d'Opticiens* (Paris : Ed. Revue d'Optique), contains a list of the works of this author between 1935 and 1948.

Gabor, D., 1949, *Proc. Roy. Soc.* A, **197,** 454; 1951, *Proc. Phys. Soc.* B, **64,** 244.

Haine, M. E., and Dyson, J., 1950, *Nature, Lond.*, **166,** 315.

Picht, J., 1931, *Optische Abbildung* (Braunschweig : Vieweg).

Rogers, G. L., 1950, *Nature, Lond.*, **166,** 237.

Some Points in the Design of Optical Levers and Amplifiers

By R. V. JONES

Department of Natural Philosophy, University of Aberdeen

Read before the Physical Society on 19th February 1949, and in revised form before the Optical Group on 9th June 1950; MS. received 21st September 1950

ABSTRACT. The design of optical levers with photoelectric amplifiers is examined and sources of instability are traced. Convection currents causing refraction variations are found in some circumstances to affect the light balance in the optical lever; they may be minimized by properly designed shrouds around the optical path. The better known causes of instability are discussed, and the appropriate remedies are described. The discussion is extended to the stability of galvanometers. The design of an optical lever and amplifier is given which can detect a change of 10^{-9} radian in a mirror 5 mm. in diameter with an indication time of 1 second and with a long-term drift of the order of 10^{-8} radian/hour.

§1. INTRODUCTION

In this paper the author examines the design of optical levers and amplifiers capable of detecting an angular change of 10^{-9} radian in the orientation of a mirror 5 mm. in diameter, with a time of indication of 1 second and stable to 5×10^{-8} radian/hour or better. Such amplifiers can be applied particularly to the measurement of the thermal fluctuations of a galvanometer, but they may also be employed with advantage in other fields.

The fundamental principle, due to Wilson and Epps (1920), is to reflect a beam of light from the mirror so that it falls equally on two sensitive receivers connected in opposition to a secondary indicator. At equilibrium the net current is zero; when the reflected beam moves, one receiver gains light at the expense of the other, and a current flows in the external (secondary) circuit. The current can be made to deflect the secondary indicator (e.g. mirror galvanometer or pen recorder) through many times the distance moved by the light beam across the receivers.

The magnification depends on the amount of light transferred from one receiver to the other for a given deflection of the beam, and on the sensitivity of the secondary system. If the device is to be used as more than a null indicator, the magnification must be kept constant. One way of achieving this is to ensure that each of the major components such as the lamp and its power supply, or the secondary system, is separately stable. Alternatively, negative feedback from the secondary to the primary system (or to the lamp itself) can be used to stabilize the amplification by relying not upon stable supplies, but on the constancy of the feedback arrangements. The first way has been chosen here largely because feedback does nothing to help against some important sources of zero instability, such as mechanical and thermal displacements in the optical system. When dealing with these problems it is necessary to stabilize several parts of the system individually, and little further effort is required to stabilize the remainder. When all reasonable precautions have thus been taken, negative feedback can be added to enhance the stability already achieved: this is not often necessary, but feedback can also be useful in varying the equivalent 'stiffness' of the deflecting system and in improving linearity.

In the present investigation an attempt has been made to find in particular the causes of instability in zero and in magnification, and hence to identify the main factors in the design of a stable optical lever with a high light-transfer in the primary system. The secondary circuit has received rather less attention, and has been made as simple as possible consistent with sensitivity. Iron–selenium photocells have been used, feeding directly a mirror galvanometer; the simplicity of this arrangement was intended to eliminate the instability that may occur in electronic D.C. amplifiers working from emission-type photocells. Such amplifiers are, of course, widely used, and they are under investigation here, but the present work has been concentrated as far as possible on the stability of the optical lever itself. The complication of an electronic amplifier would have made it more difficult to isolate the sources of instability in the lever proper. These will be described in the course of the paper, and are general to any optical lever amplifier, irrespective of the method of secondary indication. Iron–selenium cells have their own sources of instability, which will be mentioned where they have been found to affect the over-all stability of the instrument. The modifications necessary for using emission-type photocells will be outlined.

§2. PRINCIPLES

Elementary Optical Systems

Two variants of the conventional optical lever have been used in the present investigation; the first is shown in Figure 1. The source S, the lens L, and the mirror M_2 are so arranged that the image of S through L exactly fills M_2; this ensures maximum illumination of M_2 for any given selection of S, L and M_2. A diaphragm G is placed between S and L so that its image through L, and after reflection at M_2, falls in the plane of an obstacle at H; the receivers R_1 and R_2 are placed symmetrically behind the obstacle. In the most elementary design, G can be a circular hole in the diaphragm and H a circular obstacle which is slightly smaller than the image of G. In the balanced position a symmetrical annulus of light therefore passes G, and one half falls on R_1, the other on R_2. If M_2 rotates, the annulus passing H is no longer symmetrical, and one receiver gains light at the expense of the other. It is not essential that there be any obstacle at H, because it does not affect the net transfer of light from R_1 to R_2. It does, however, serve to block out the central light, which would otherwise generate a large but useless signal in each receiver. Hill (1948) used such an obstacle to reduce this signal in order to cut down the steady current (and hence the random fluctuations in this current) from each receiver. The obstacle, however, is of even greater use, for it reduces the effect of any accidental variation in any other component of the system (the source brightness, for example) on the output of the receivers. Ideally, the width of the annulus of light allowed to pass H should not be much greater than that of the largest deflection which it is desired to measure.

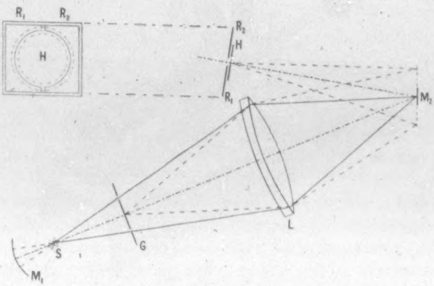


Figure 1. Elementary optical system.

It follows that the unused light can be blocked out at the start by making G an annulus also; the sensitivity of the system will be unaffected. The logical extension of this idea is that G should become a family of concentric annuli and H a similar family, each member of H being slightly smaller than the image of the corresponding member of G. This system is more sensitive than that using single annuli, since the wasted light inside the original annulus is now utilized; the overall sensitivity is proportional to the sum of the diameters of the annuli.

Rectilinear grids are more efficient than annuli, but such grids cannot be used in this simple system, since field curvature and pincushion distortion are both serious; they are caused by M_2 acting as an optical stop a long way to the image side of L. This fact limits the usable shapes to those which preserve their form in the presence of distortion, and which can be mechanically generated. These are annuli, already considered, and radial slots: both patterns are also unaffected by astigmatism. Systems with four concentric annuli have been tried; owing to the field curvature, the annuli at G have to be staggered along the optic axis, the innermost being nearest S, in order to produce a plane image at H. The annuli at H are then fitted to the resulting image. The gain in sensitivity over a single annulus is three times.

A system with eight annuli has been tried, but the individual rings are troublesome to mount, and the same magnification can be obtained more easily by using radial slots.

Diaphragms and obstacles with forty radial slots have been tried; these are easier to construct, since they can be machined without hand-fitting. Aberrations, however, spoil the image towards its edges, and the overall gain in sensitivity over a single annulus is not more than six times. These systems are optically clumsy compared with that to be considered later, but they are very satisfactory for magnifications up to about 40,000 (measured by the ratio of linear movement of secondary indicator to that of primary beam across photocells) when used with iron–selenium cells; they have the merit of requiring the fewest possible components to be held rigidly. A drawing of a complete lever system to this design is shown in Figure 2. The mechanical details are much the same as those of the system to be described later. One optical detail, however, may be worth mentioning immediately. It is important that the only thing to upset the light-balance in the receivers should be a rotation of the mirror M_2; all other components must therefore be held as rigidly as possible. The filament of the source S cannot be as rigid as the other components, but the system can be stabilized against small movements of S by placing a concave mirror M_1, centred approximately on S, behind it. The mirror forms an inverted image of S alongside it, and if S moves off axis in any direction, the image moves correspondingly in the opposite direction; the illumination of G (which determines the light-balance in the final image of G at H) therefore tends to remain balanced. At the same time the amount of light falling on the receivers is increased; measurements show that this increase is about 60%, while the disturbance of light-balance for a given displacement of the source is reduced to about 40% of its former value. The proportionate increase in stability is therefore about four times.

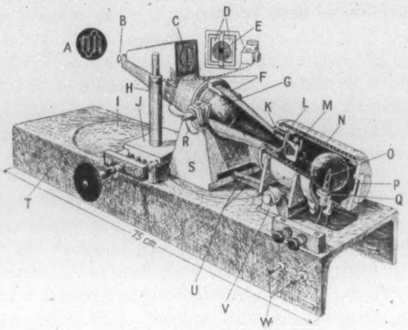


Figure 2. Optical lever using elementary system.

A. Light pattern on galvanometer mirror.
B. Galvanometer mirror.
C. Image of ring system.
D. Two 22 × 40 mm. cells mounted side by side.
E. Ring obstacle.
F. Metal shrouds for optical path.
G. Condenser and image-forming lens.
H. Post holding photocells.
I. Scraped table for galvanometer.
J. Kinematic slide.
K. Clearance between shroud and lamp-house to permit relative adjustment.
L. 5 mm. ON 20 heat filter.
M. Staggered ring system.
N. Water-cooling coil.
O. 12 v. 48 w. class F lamp.
P. Concave mirror in focusing mount.
Q. Copper sleeve soldered to lamp-cap, sliding in outer sleeve fixed to side-tube of lamp-house.
R. Wing nut locking lens ring to trunnion (one side only).
S. Lens trunnion (built up on four sides for rigidity).
T. 8 in. inverted-U mild steel girder.
U. Locking bars for trunnions.
V. Lamp-house trunnion.
W. Cooling connections.

Improved Optical System

A simple modification (Figure 3) of the foregoing optical system, which incidentally brings it more into line with conventional systems, makes it possible to employ multi-element grids, with a corresponding improvement in sensitivity.

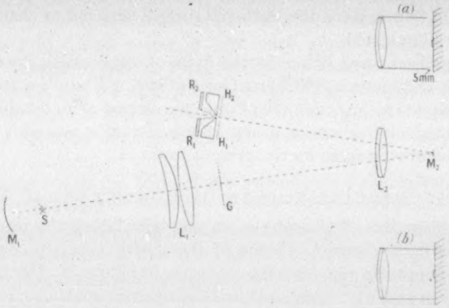


Figure 3. Autocollimating optical system. (*a*) and (*b*) are alternative positions of sheath referred to in the test for convection currents illustrated in Figure 5.

As before, a lens L_1 focuses the image of S on to M_2. A grid G is placed between L_1 and M_2, and lens L_2 placed in front of M_2. An image of G of exactly the same size as G is formed in the plane H_1H_2 by light passing twice through L_2, with reflection by M_2. This arrangement is selected for several reasons. First, the unit magnification makes it possible to use two replica grids, one at G and the other at H_1H_2. Secondly, the lens L_2 and mirror M_2 form a system in which M_2 acts as a small and symmetrically placed stop between L_2 and its image. Distortion is therefore eliminated and a rectilinear grid will have a rectilinear image. Coma is also eliminated. This is in fact the standard autocollimating position, and focusing is not dependent on the position of M_2. It is an advantage to have a distance of at least 2 cm. between L_2 and M_2 so that L_2 can be clear of any projections, such as a galvanometer magnet, around M_2. The difficulty in providing the necessary clearance lies in the fact that the further L_2 is from M_2 the larger the section of the cone of light converging to M_2 which is cut by the surfaces of L_2. This means that the aberrations in L_2 become more serious and set a limit to the definition of the image; nevertheless a working distance of 4 cm. has in fact been achieved, with sufficient definition to use grids 4 cm. square of about 66 lines/cm. with a lens of 15 cm. focus.

The second grid H_1H_2 is split and mounted in two halves, left and right, so that the lines and spaces in one half are exactly out of phase with those in the other. If, for example, the spaces of the image of G coincide with those of H_1, then the latter will allow all the light in its half of the image to pass it; at the same time H_2 will obscure all the light in its half. A movement of the image of half a grid spacing relative to H_1 and H_2 will completely reverse the position, all the available light passing H_2 and being obscured by H_1; at an intermediate position each side lets through half the maximum light and the system balances. The resultant magnification is therefore proportional to the number of lines in the grid; the useful range of movement is, of course, inversely proportional to this number.

The use of relatively fine grids reduces trouble caused by point-to-point variations in sensitivity over the surfaces of the photocells. The cells can be mounted at a sufficient distance (1 to 2 cm.) behind the image grid for the individual grid lines to become so inter-diffused that the light on each photosurface appears uniform, and the effect of rotating the optical lever is only to change the intensity of illumination over each cell surface, and not to change the portion of the surface illuminated.

Grids have been used before in this type of application, notably by Pfund (1929), Barnes and Matossi (1932) and Golay (1949). Pfund's system anticipated the present one in the principle of splitting the second grid, but his paper seems to have attracted so little notice in the literature that it was only found during the preparation of references for the present paper.

§3. SPECIAL POINTS OF CONSTRUCTION

Base and trunnions. A drawing of an amplifier using the improved optical system is shown in Figure 4. Some of the details that are common to both types of amplifier can be seen from the other side in Figure 2. The base is a length of 8 in. U-sectioned mild steel girder. The top surface is machined and scraped flat over those areas which carry components of the system. The major machining is done some months before the amplifier is completed, so that any subsequent distortion (which sometimes occurs) is removed in the final scraping. The under surface of the trunnion supporting the lamp-house (and also that supporting the

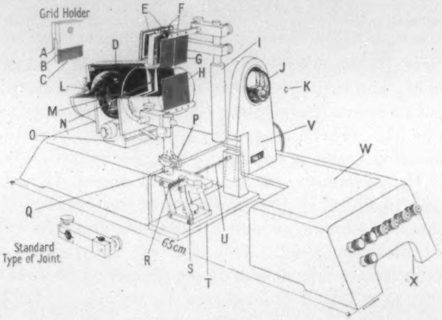


Figure 4.

A. Paper.
B. Glass.
C. Aluminium grid on glass.
D. Cooling tubes.
E. 22 × 40 mm. photocells.
F. Prisms refracting light on to cells.
G. Split grid.
H. Object grid.
I. Photocell column (drilled to take cell leads to base).
J. Autocollimating lens in screw mount.
K. Mirror.
L. Concave mirror
M. ON 20 heat filter.
N. Condenser.
O. Trunnion.
P. Locking collar for object grid adjustment.
Q. Pin carried by spring movement projecting through slot in base.
R. Adjustable bearing for wobble pin.
S. Parallel spring motion.
T. Bridge to carry spring motion.
U. Wobble pin.
V. Lens pillar locked to base by single screw and supporting front shroud (not shown).
W. Scraped platform.
X. One corner of base machined off and replaced by adjustable foot (not shown) to be screwed down to make fourth contact with table.

lens in the older system) is scraped slightly concave to give contact at the corners of the base of the trunnion. The trunnion is locked to the base by a single substantial screw pulling into a nut on the underside of the girder: this gives firm locking, and at the same time allows a choice of positions and orientations set by the size of the slot in the base through which the locking screw passes.

Lamp-house. The lamp-house consists of a brass tube of diameter 3 in. and wall thickness $\frac{1}{8}$ in. The source, a 12 v. 48 w. class F projection lamp, has a filament of suitable shape for roughly filling a square or circular mirror; it is operated at 9·5 v. and 3·5 amp. for longevity. The ordinary screwed contacts have been found unreliable as the lamp warms up, and power is therefore supplied through a lead soldered directly to the centre contact of the lamp and through a copper sleeve soldered over the screw thread (Figure 2). The sleeve is held in metallic contact with the body of the instrument, which provides the return path for the lamp-current.

Lens mounting. In earlier models the condensing lens (an achromat 10·5 cm. focal length, 7 cm. aperture) was mounted in a heavy brass ring, held in a rigidly built trunnion (Figure 2), and locked on one side only, allowing the axle on the other side to rest freely in its bearing on the trunnion. In later models (Figure 4) the condenser assembly (now of conventional design, 9 cm. focal length and 6 cm. aperture) is mounted directly on the lamp-house without serious loss in stability.

The autocollimating lens (an achromat 13·5 cm. focal length, 4 cm. aperture) was originally mounted at the end of a box-sectioned extension from the ring holding the condenser; this was not stable enough, and it proved necessary to hold the lens, and indeed most other components, as directly as possible from the basic girder as shown in Figure 4. Coarse focusing, and setting the lens square to the system, can be done by sliding and rotating the lens pillar under the single fixing screw which locks it to the base; fine focusing is achieved by rotating the lens in its screw mount. High stability of the autocollimating lens is essential, since any transverse movement of the lens moves the image of the grid twice as far; a stability to at least 10^{-7} cm. over an hour is therefore required.

Heat filter. The lamp-house also includes a heat filter of ON 20 glass, 5 mm. thick, to cut out infra-red radiation which is not detected directly by the photocells, but which contributes, as pointed out by Preston (1946), to their fatigue, and which might distort the mirror by excessive heat.

Grid construction. A master grid of wire is wound, and replicas made by evaporating aluminium *in vacuo* on to glass plates placed in contact with the master. One replica is used at G, another is cut into two pieces to form the split grid H_1H_2.

In making the grids it has been found advisable to use a wire/space ratio of about 45/55, so that on the replicas the opaque/transparent space ratio is about 55/45. This gives a little latitude in the adjustment of the image of the first grid relative to the split grid, if maximum obscuration is desired. With grids of 66 lines/cm. a ratio of 12 : 1 has been obtained between the light passed in the most favourable and most unfavourable positions of image relative to second grid. This indicates that the image definition is adequate.

Grid mountings. The object and image grids must be mounted so that their positions are stable to at least 10^{-7} cm. over an hour, and yet so that their relative transverse positions are accurately adjustable. The required adjustment is obtained by using either a semi-kinematic or a parallel spring slide (§4). The grids themselves can be held sufficiently firmly by gripping them on one edge only to a depth of about 1 mm. between two suitably milled strips of mild steel (Figure 4), one side being packed with a compressible lining such as paper. The grids are then attached either to the base or to the adjustable slide by an assembly of sliding and rotating (but lockable) joints of appropriate degrees of freedom.

In one design the object grid was mounted directly on the front of the lamp-house in an attempt to economize in supports and adjustments. This failed, owing to the expansion and contraction of the lamp-house walls with mains water temperature, and to the flexure of the lamp-house under mechanical reaction to the water flow. The latter was sufficient to move the secondary galvanometer spot by several centimetres, even though the water flow was moderate and the lamp-house substantial.

Slides. Either the object grid must be mounted directly on the base, and the image grid on the adjustable slide, or vice versa. In earlier designs the image grid was the one to move, but in the most recent design the object grid is moved.

The semi-kinematic type of slide moves on two hardened steel rods which are ground and honed and then mounted parallel to one another on short brass pillars rising from the base. The slide itself is of brass, with two short vees and one flat milled on its underside, in contact with the rods. It is spring-loaded against a micrometer head. For moderate amplifications this type of slide is satisfactory, since it has only to be adjustable to the nearest micron or so. At higher magnifications, where closer adjustment is required, it is not entirely satisfactory, since friction between slide and rods is variable: the slide then tends to stick and to slew across the rods. This fault, although microscopic, can be shown by the fact that the secondary galvanometer spot sometimes does not move when the micrometer is turned, or moves for a short distance in the contrary direction to that in which it will ultimately go.

The parallel-spring slide shown in Figure 4 is completely free from the defects of the semi-kinematic type, when properly made. The details of design are discussed elsewhere (Jones 1951). An optical micrometer (a rotatable refracting plate) could be interposed in the optical system for finer adjustment but has proved unnecessary.

Shrouds and convection currents. It is important to provide shrouds enveloping the optical path as closely and as completely as possible. They obviously prevent dust particles from falling through the beam, thus upsetting the light balance, but they also reduce disturbances from a more insidious cause—convection currents of air in the light-path. A system such as the present one is at least as sensitive as a schlieren system in detecting changes in refraction in the two balanced halves of the optical path, and the changes in light-balance due to this cause gave much trouble until they were identified and eliminated. They were identified first by accentuating them with coal gas and ether vapour, and then by eliminating them by complete evacuation of a specially designed system. Fortunately this system was also tested with air inside it, and it was found to be free from convection troubles. The shield around the path between lens, mirror and photocells was then removed and the trouble reappeared even though the whole apparatus was still under a dust cover. This test led to the discovery that convection troubles could be sufficiently reduced for most purposes (by a factor of fifty in some cases) by shrouding the optical path with a container of minimal dimensions. Part of the improvement may be due to the uniform temperature around the path provided by the metal shroud; this shroud may also stabilize such currents as do exist inside it, so that these do not waver about as they do in larger containers.

Without the shrouds the convection disturbances in some cases (which were observed when the mirror at M_2 was fixed) exceeded by a factor of three those due to the Brownian motion of the galvanometer to which the optical system was to be applied. Convection currents have been mentioned by other workers as causing undesirable motion of galvanometers and electrometers by virtue of the draughts that they create around the suspension, but I have not seen any previous reference to their effect in upsetting the light balance, which can only be shown up by using a firmly fixed mirror. It is possible that some of the unsteadiness of galvanometers in excess of Brownian movement which has in the past been attributed to vibrational disturbances has in fact been due to refraction variations. Ornstein and his co-workers (1927) suspected such variations, but considered that their effect was immaterial in their investigations.

Figure 5 (Plate)* shows a typical effect of convection currents in disturbing the light balance of the amplifier sketched in Figure 4. In this instance the mirror was fixed. Figure 5 (*b*) shows the secondary trace when a cylindrical sheath extended from around the lens L_2 right up to the surface of the mirror M_2, as in Figure 3 (*b*), completely sealing the air space between L_2 and M_2. Figure 5 (*a*) shows the corresponding trace when the sheath stopped 5 mm. away from M_2, as in Figure 3 (*a*), leaving the space between L_2 and M_2 partially open to the rest of the air under the main dust cover. All other conditions were identical for the two traces.

§ 4. OPERATION

Power supplies. Large capacity (150 ampere-hour) submarine batteries have proved satisfactory for supplying the lamp current.

Several other methods of supply have been tried, but nearly all have failed. For constancy in light output to within 0·5% the lamp voltage must be constant at least to within 0·2%. Few stabilizing systems achieve this figure over a period of several hours, but an electronic stabilizer, developed by Mr. J. C. S. Richards, has proved satisfactory. It makes use of a saturable choke, and takes as reference the thermionic emission variation with heating current of a tungsten filament. It delivers up to 10 amperes at 12 volts and, over an hour's run, shows a total voltage variation of not more than 0·1%. Over twelve hours the total variation may be 0·4%, which is better than most batteries. Since this stabilizer has a wide application in lighting stabilized lamps, a full description may be published separately.

Troubles peculiar to iron–selenium cells. Apart from troubles such as convection currents, release of mechanical strains, and thermal drifts, which are common to all types of optical lever, there are some that are peculiar to iron–selenium cells; these must be made negligible.

The least troublesome, but most mentioned, effect is that of fatigue: it can be made small, as Preston (1946) points out, by use of an infra-red excluding filter. Moreover, when the highest stability is sought, the system must run with the lamp lit for several hours before observations are made, and this gives the photocells ample time to reach their ultimate-fatigue condition.

---

* For Plates see end of issue.

A more troublesome effect arises because pairs of cells that are matched to give equal outputs at a particular light intensity may not respond equally to small changes in that intensity, particularly if the colour balance of the light changes at the same time (Preston 1950). This means that while the system is balanced for a particular lamp voltage, it may go out of balance if this voltage changes; and the departure from balance may cause a serious change in the reading of the secondary galvanometer. The defect can be much reduced (twenty times in a trial case) by placing a suitable resistance in series with the cell that has the greater logarithmic intensity to changes in light intensity, and by adjustment of the division of light between the cells. Unfortunately, this solution is not always satisfactory, for it sometimes happens that a cell which gives a steady output when feeding, for example, a 350-ohm galvanometer directly, becomes 'noisy' when balanced by a resistance of a few hundred ohms in the manner suggested above. This behaviour is unexplained, but it has happened with two different pairs of cells. The amount of noise developed is not large, being only equivalent to the order of 10^{-7} radian in the levers used, but it has to be eliminated for highest sensitivity. The solution has been to ask the makers to provide cells matched as well as possible for the illumination (0·05 to 0·15 lumen/cm^2) at which they have to operate.

Iron–selenium photocells have high temperature coefficients which vary from cell to cell. This has probably been the worst cause of thermal instability in the whole system, exceeding considerably the troubles due to thermal expansion in the base and in the components (except for one case where the object grid was held directly on the water-cooled lamp housing). One cure is to select cells which are matched also in this respect: most cells are improved when operating into a low-resistance circuit because the temperature coefficient then decreases.

Temperature control. Another method of dealing with temperature effects in the photocells, and in other parts of the amplifier, is to 'thermostat' the apparatus; when this is done it is advisable also to run the cooling water from a constant head. The lever is provided with a 'Perspex' outer case to exclude draughts and to reduce temperature fluctuations; when the thermostat control (a Sunvic TR1) is sited on the case, and 3 kw. of heaters arranged roughly around the edges of the floor (average distance four feet from the concrete pillar supporting the amplifier), the temperature is confined within a range of 0·05° c. The fluctuation of the zero due to this residual change in temperature is then equivalent to about 10^{-8} radian with a period depending on the thermostat cycle. All harmful expansion effects can be minimized by using the same material (mild steel) as far as possible throughout the apparatus. When this is undesirable it is generally easy to arrange that expansion will not seriously displace the major components relative to the optic axis.

Operation of pen recorders. With iron–selenium cells it is convenient to use photographic recording of the trace from the secondary galvanometer, but it is desirable for many purposes to dispense with photography and to operate a pen recorder. Although such a recorder requires much more power than the iron–selenium photocells can give, two alternatives are readily available. The first is to add a magnetic amplifier to the photocells, the second to change to emission photocells and amplify their output thermionically. Although the former method has not been tried, the available literature on commercial magnetic amplifiers suggests that they are sufficiently constant in amplification and zero stability to merit trial.

The second method has been tried by substituting a Cintel type VA40 split photocell for the pair of 22 × 33 mm. iron–selenium cells. With a 35-mm. square grid of 28·6 lines/cm., a 1-micron ($2{\cdot}5 \times 10^{-6}$ radian) displacement gives a signal of 50 mv. when using 1MΩ load resistors. Since a D.C. amplifier can be designed fairly easily to drift less than 1 mv/hour, the equivalent angular drift added by such an amplifier is less than 5×10^{-8} radian/hour. This, although detectable, is not serious for many purposes. Apart from making possible the use of pen recorders, the extra amplification can also be used to reduce the response time of the secondary system. Hill (1948), for example, has used a cathode-ray tube as secondary indicator, with negligible response time.

In order to avoid the drifts involved in D.C. amplifiers, the light beam in the primary system can be chopped and an A.C. thermionic amplifier used with the photocells. This has been done by Milatz and Bloembergen (1944). Alternatively, the primary mirror can be oscillated (e.g. by interrupting the radiation falling on a thermopile to which the primary galvanometer may be connected), and the consequent alternating current from the photocell amplified. With all these possibilities the problem of secondary indication can be solved for any reasonable set of conditions.

Miscellaneous precautions. The optical lever system itself is relatively insensitive to vibration, but trouble may arise when the fixed mirror is replaced by one on a suspension, such as that of a galvanometer. It has not been necessary during the present investigation to take any more precautions against vibration than to place the whole system on a concrete pillar built directly into the soil below the laboratory.

It is possible to use a primary galvanometer of 2 seconds period in place of the fixed mirror, even though relatively heavy machinery (e.g. a three-ton shaper) is operating within thirty feet of the amplifier. The effect of the workshop vibration in the case of a critically damped galvanometer on an average day was to increase the r.m.s. fluctuation from 14·8 mm. to 15·8 mm. The former figure can be entirely attributed to thermal fluctuations, so that the increase is negligible for almost all purposes. If the suspension is but lightly damped, workshop vibration adds greatly to its fluctuation. In this case it is necessary to work out of workshop hours, when seismic and other disturbances do not contribute more than 10% above the theoretically expected thermal fluctuations.

If further anti-vibration precautions should be necessary at some other site, the vibration insulation system should not include 'expanded' rubber. This has been found to collapse in small jerks for days after the load has been placed upon it, and to disturb any suspended system which it may be supporting.

The mirror on which the light is concentrated may sometimes flex under the heat which is absorbed; this has not been found a serious matter, although it has been detected, because the mirror usually takes up a definite figure for any one value of the lamp brightness, and this of course is kept constant for other reasons.

Changes in level of the building may distort the pillar on which the apparatus is mounted sufficiently to cause an observable shift of zero. An observer, for example, is detectable as he walks around the room by the change in level of the 'solid' floor beneath him. This is equivalent to a deflection of about 5×10^{-9} radian. A similar result can be achieved by intentionally 'bending' the concrete pillar (which presents a 1 m. cube above ground) with one finger; the bending is transmitted to the base of the instrument and the alignment of the optical components disturbed. This effect illustrates the force of Rowland's instruction to regard any instrument in which mechanical rigidity is required as being made of jelly. Changes in floor level may alter the zero of any suspended system which is not perfectly balanced; this effect is usually more serious than that due to the upsetting of the optical alignment.

In view of the many possible sources of trouble, it may be helpful to record that no trouble has arisen in three years from possible fluctuations in lamp brightness due to the movable contact of the rheostat controlling the current through it. Trouble certainly arises if the lamp contacts themselves are not soldered, but the rheostat contact has given no trouble.

Primary galvanometer precautions. Three Tinsley type 4500 galvanometers were tested, and two of them had zero constancy within the r.m.s. of Brownian fluctuation for periods of many hours. The only precautions which proved necessary (although many were tried) were (i) to keep the room temperature steady, (ii) to keep the instrument level, (iii) to mount the galvanometer suspension and magnet in a small airtight box, (iv) to mount this box on a platform which was prevented from 'creeping' about the base of the amplifier, by the use of plasticene if necessary, (v) to avoid inductive loops in the primary circuit, (vi) to make the circuit as compact as possible, and to avoid, if possible, sources of thermo-E.M.F's, (vii) to 'earth' the system properly.

The galvanometers were operated for some months in evacuated containers to avoid the dangers of convection currents disturbing the suspensions, but this precaution proved unnecessary so long as the external containers were small and airtight. A prolonged test, amounting to about 8,000 observations, was made to compare the zero fluctuations of the same galvanometer at atmospheric pressure and *in vacuo*; it was found that disturbances due to convection currents around the suspension were insufficient to affect either the apparent r.m.s. of the thermal fluctuations or the long-term drifts of the system. Had they increased the former quantity by 5% or the latter by 50% their effect would have been recognized. It may be of interest to recall that both Cavendish (1798) and Boys (1895) noticed that convection currents disturbed their suspensions, and that Boys found that these currents could be minimized by making his apparatus and container small. The present investigation into refraction variations in the optical path has confirmed Boys' finding; the apparently contradictory observation that convection currents do not upset the suspension of the galvanometer is due to the fact that, as in many commercial galvanometers, the walls of the suspension tube are sufficiently near to the suspension itself to make convection currents negligible, especially when this tube is enclosed inside a small airtight box.

In general, zero stability was good, and even after relatively large deflections (4×10^{-4} radian) the galvanometers returned to their original positions well within the r.m.s. of their Brownian movement. Continuous records were taken showing no serious variation in zero over periods of 20 hours (Figure 7, Plate).

Linearity. It is difficult to check the linearity of the amplifier at very low inputs, but departure from linearity could clearly become serious when the deflection of the primary beam exceeds more than about one-quarter of the grid spacing. This has been confirmed by reducing the sensitivity of the secondary circuit and then applying measured deflections in 10-micron increments to the grid and photocell assembly by means of the micrometer adjustment. With a grid of 28·6 lines/cm. the useful (i.e. apparently linear) range of deflection was about 80 microns. The change in current for this deflection amounted to 14 μa., and at no point did the deflection output curve depart from a straight line by more than 0·1 μa.; the residual error may have been largely due to the difficulty of setting the micrometer.

Grids with 66 lines/cm. gave a linear range of about 40 microns. Such a range is, of course, far beyond that which the secondary system is capable of indicating when operating at highest sensitivity, since it implies a deflection of 60 metres in the secondary galvanometer; but it shows that within the limited range which the latter can measure the linearity of the amplifier is likely to be good.

The amplifier with a 28·6 lines/cm. grid was tested with a primary galvanometer in position; the amplification was about 100,000, and Brownian motion showed an r.m.s. of the order of 10 mm. in the secondary. Currents in the ratio $\pm 1 : \pm 2 : \pm 3$ were then passed through the primary galvanometer, the deflection for the smallest currents being ±142 mm., as averaged over 1 minute's observation for each reading. The corresponding figures for twice and three times the smallest currents (when corrected for geometrical distortion of large deflections due to a flat scale in the secondary system and for non-linearity in the secondary galvanometer) were ±283 mm. and ±424 mm. The maximum departure from linearity over this range was therefore not more than about 2 mm., which is about the probable error due to the Brownian motion when averaged over a minute.

Tests, both by mechanical and by electrical displacements, of the image of the first grid relative to the split grid thus indicate that linearity is substantially better than 1% over a range given by one-quarter of the grid spacing. Errors found within this range may in fact be due in part to the difficulty of accurately producing small deflections.

For a wide range of linearity coarse grids must be used or feedback suitably applied.

Constancy of amplification. Since the light-output of the lamp is roughly proportional to the 2·5th power of the lamp voltage, the amplification will be correspondingly sensitive to this source of variation. Moreover, the lamp brightness will diminish with age. Neither of these two factors has been found serious when working with large-capacity batteries and under-running the lamp. Tests with a fixed mirror have shown that the amplification does not vary during an hour by more than 0·2%.

Both constancy and linearity of amplification can of course be improved, if desired, by adding negative feedback.

§ 5. PERFORMANCE

The last model constructed had the following performance: Magnification between light movement on photocells and that of secondary galvanometer (Kipp double coil, 300 ohm winding, 1·3 second period) recording at 1·5 metres: $1{\cdot}5 \times 10^6$. 1 mm. on secondary record corresponds to $1{\cdot}5 \times 10^{-9}$ radian movement of mirror. Estimated minimum movement of mirror probably detectable within 1 second: 10^{-9} radian, corresponding to a displacement of the edge of the mirror by 2×10^{-10} cm. Changes in zero attributable to convection currents (when all precautions are taken short of evacuation): of the order of

3×10^{-9} radian, with equivalent period of about 10 seconds. Changes in zero attributable to the thermostat cycle: of the order of 10^{-8} radian, period varying from 20 minutes to 2 hours. Long-term drifts in zero due to miscellaneous causes, such as ageing of materials of building level, lamp-ageing, etc.: 10^{-7} radian in 12 hours.

A typical sample of the secondary galvanometer trace with a fixed primary mirror at a scale of 1 division to $1{\cdot}5 \times 10^{-9}$ radian is shown in Figure 5 (*b*) (Plate). A sample of the record when a primary galvanometer of period 2·1 seconds replaces the fixed mirror at an approximate scale of 1 division to $1{\cdot}5 \times 10^{-8}$ radian is shown in Figure 6 (Plate), with a fixed-mirror record at the same amplification for comparison. Figure 7 (Plate) shows the comparative freedom from overnight drift achieved with one particular galvanometer on several occasions.

ACKNOWLEDGMENTS

The construction of an evolving succession of amplifiers has been undertaken by the workshop of this Department, and several of the mechanical features are due to the workshop staff. My thanks are due to Mr. D. C. Gall of Messrs. Tinsley and to the Admiralty Research Laboratory for the loan of galvanometers, and also to Mr. Gall for drawing attention to the possibilities of grids in the amplifying system. It is a pleasure also to acknowledge the loan by Professor A. Michels, of Amsterdam, of a Kipp Double Coil Galvanometer which has been used throughout for the secondary recording.

REFERENCES

Barnes, R. B., and Matossi, F., 1932, *Z. Phys.*, **76,** 24.
Boys, C. V., 1895, *Phil. Trans. Roy. Soc.* A, **186,** 1.
Cavendish, H., 1798, *Phil. Trans. Roy. Soc.*, 469.
Golay, M., 1949, *Rev. Sci. Instrum.*, **20,** 816.
Hill, A. V., 1948, *J. Sci. Instrum.*, **25,** 225.
Jones, R. V., 1951, *J. Sci. Instrum.*, **28,** 38.
Milatz, J. M. W., and Bloembergen, N., 1944, *Physica*, **11,** 449.
Ornstein, L. S., Burger, H. C., Taylor, J., and Clarkson, W., 1927, *Proc. Roy. Soc.* A, **115,** 391.
Pfund, A. H., 1929, *Science*, **69,** 71.
Preston, J. S., 1946, *J. Sci. Instrum.*, **23,** 173; 1950, *Ibid.*, **27,** 132.
Wilson, W. H., and Epps, T. D., 1920, *Proc. Phys. Soc.*, **32,** 326.

Circulations Occurring in Acoustic Phenomena

By G. D. WEST
Physics Branch, Military College of Science, Shrivenham

MS. received 5th December 1950

ABSTRACT. An investigation is made, by means of smoke particles, of the movements of the air in the neighbourhood of a vibrating reed. At very low frequencies, the paths of the particles exhibit a pattern which, at first sight, resembles the lines of hydrodynamic flow. A more careful examination at sufficiently high magnification shows, however, that the particles do not move along the hydrodynamic lines but trace out small ellipses.

In addition, a mass circulation of the particles is superimposed on the above motion. At low frequencies this circulation is feeble, but with increasing frequency it strengthens and soon becomes the salient feature. Eventually yet another circulation is superimposed.

It is shown how this work connects up with that of other observers, and emphasis is placed on the common occurrence, in many acoustic phenomena, of circulatory motion in association with vibratory motion.

§ 1. INTRODUCTION

THE present work arose from the observation (West 1946) of the way in which smoke-laden air streamed from the end of an excited resonator. Although the phenomenon at first appeared simple, it was soon found to involve complicated motions (Andrade 1947, Lawson 1947, Simpson 1947, West 1947) of a circulatory kind. Certain types of such circulations have already been described (Andrade 1931, Carrière 1931, 1940). The present investigation makes a further study of related types—particularly those which occur at low frequencies.

Method Employed

In the initial experiments the arrangement used was similar to that shown in Figure 1. A piece of watch-spring A, vibrates in a box filled with tobacco smoke illuminated laterally by a strong beam of light B. By means of a microscope C, the smoke particles can be seen as bright points on a dark ground. For continuous observations it is convenient to maintain the spring in vibration by an external electromagnet excited by the 50 c/s. alternating current supply mains, but for very low frequencies (say 3 or 4 vibrations per second) the spring can be weighted at the top and it then goes on vibrating for a considerable time without external aid. For frequencies of some hundreds of vibrations per second, the strip of watch-spring was mounted on the vibrating element of a reed-telephone. Only, of course, at resonant frequency was a large amplitude possible.

§ 2. RESULTS

With a low frequency of the order 3 or 4 per second, a pattern similar to that shown in Figure 2 (Plate*) is seen. It will be observed that there is a general resemblance to the well-known hydrodynamic flow shown in Figure 3 which takes place when a lamina moves forward in a perfect fluid. In the case of vibratory motion of a lamina in such a fluid, however, each particle moves backwards and forwards over a short arc of stream line but, in an actual fluid, as

---

[*] For Plates see end of issue.

can be seen from the photograph, most particles trace out small oval curves. Similar curves are traced out by small particles if the air in the box is replaced by water or even by a mixture of water and glycerine. There is, however, always superimposed on the above motions a general streaming of the fluid. This consists, as shown in an actual photograph (Figure 4, Plate), of four circulations. The smoke moves towards the front and back of the vibrator and away at the sides*. At frequencies such as those shown in Figure 2 the streaming, although present, is feeble, but if the frequency or amplitude is increased, the streaming increases also. Experiments were carried out in air and water. In air the circulation is reduced in speed by reducing the pressure and, since the viscosity is constant over a wide range, this result must be attributed to a change in density. In water the viscosity can be changed without making much difference to the density by the addition of glycerine. The result is once more to reduce the circulation.

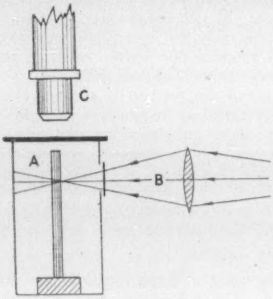


Figure 1.

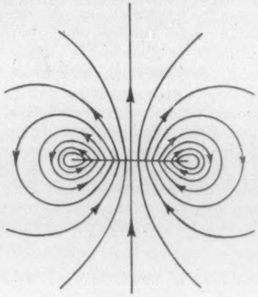


Figure 3.

With large amplitudes or high frequencies it was found that more complicated circulations made their appearance and that they were not restricted to a horizontal plane. When, however, the watch-spring strip was replaced by a cylindrical pin, horizontal circulations were maintained even at high frequencies and amplitudes. Thus, at a frequency of 550 vibrations per second and with an amplitude of less than 0·040 cm., the circulation around a pin of diameter 0·077 cm. was very similar to that shown in Figure 4. With an increase in amplitude to 0·045 cm., however, this circulation had shrunk to a region near to the cylindrical surface and around it was a vigorous circulation in the opposite sense. The general appearance is represented in Figure 5. At a slightly greater amplitude, it was hard to observe any inner circulation at all. The new circulation takes place, according to the above data, when Reynolds' number is more than about 40. In so far, therefore, as its appearance is somewhat catastrophic, it is in marked contrast with the circulation originally described.

In the previous experiments it is the reed that moves and the air that remains stationary. It is easy, of course, to arrange experiments in which the reverse is true. This was done by fixing a telephone at the end of the long thick-walled brass tube indicated in Figure 6.

---

* The streaky appearance which develops in smoke used as an indicator is considered by Andrade (1931) in his paper.

For very low frequencies of 6 or 7 per second the telephone was replaced by a piston, set in reciprocating motion by a small motor.

There is no difficulty, of course, in anticipating what should be expected in a perfect fluid. It is only necessary in fact to apply a transformation to Figure 3 such that the reed is reduced to rest and the air set in motion. Figure 3 thus transforms into the well-known diagram of the flow of a perfect fluid past a lamina.

The actual smoke particles form a pattern of this type provided that the amplitude is small. The ovals in Figure 2 appear again, however, but much elongated and more or less along the flow lines. This would also be anticipated from a simple geometric transformation of the ovals in Figure 2.

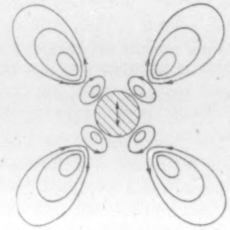


Figure 5.

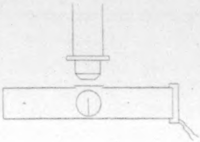


Figure 6.

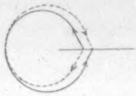


Figure 7.

With increasing frequency or amplitude, a circulation similar to that in Figure 4 gradually appears. Likewise, at still higher amplitudes or frequencies, the second circulation shown in Figure 5 makes its appearance and indeed seems to take precedence particularly if the reed is replaced by a small cylinder. This is essentially the circulation described by Andrade (1931) and is, of course, confirmed by the present experiments. Andrade also refers to an 'ill defined inner circulation'. Doubtless this is the first circulation described, which is difficult to see at high frequencies but which is very much in evidence at low frequencies.

§ 3. EXPLANATION OF RESULTS

It is possible to form some idea of the way in which the oval paths of the smoke particles arise if we make use of the fact (which can be demonstrated easily be a simple experiment) that, in an actual fluid, the stream lines such as those shown in Figure 3 are not symmetrical about the lamina around which the flow takes place. In Figure 7 therefore, are shown diagrammatically two stream lines that might refer to an actual fluid.

The full line refers to the motion of the strip upwards and the broken line to that downwards. A smoke particle will thus find itself moving along a small arc of the full line when the upward stroke is at its maximum velocity and along the broken line when the downward stroke is at its maximum velocity. At the extreme limits of the stroke the particle would be half way between the two curves and would be moving with its maximum speed at right angles to them. Such motion could be regarded as approximately simple harmonic and one quarter period out of phase with the simple harmonic motion along the curves.

The resultant motion is thus an ellipse, and the paths of the particles, as shown in Figure 2, can thus be explained.

To confirm this, a stroboscopic disc was constructed in which holes, equivalent in circumferential length to one half period, were covered half with red gelatine and half with green. The field was thus first illuminated by one colour and then by the other. When the disc was run synchronously, the directions of motion of the particles could thus be determined and were in entire agreement with what would be predicted by the curves shown in Figure 7.

The explanation here given of the results—depending as it does on the difference between up-stream and down-stream motion—is in agreement with that given by Carrière (1940), although his experimental conditions are different.

It is possible on the basis of asymmetric flow to give a very general explanation of the way in which the circulations shown in Figure 4 arise. A moving reed or cylinder appears to drag more fluid in a direct line behind it than it pushes forward, and the superimposition of the two flows, such as may be imagined in oscillatory motion, leads to a net flow on either side of the strip towards the centre. On reaching the centre the fluid moves laterally outwards and eventually returns again in the circulation. The excess fluid which moves to the front and back of the vibrator and which then moves out at the sides, suggests that some imitation might perhaps be given by means of source and sink flow. Let us therefore suppose the vibrating cylinder replaced by a stationary cylinder with a line source parallel to the axis on each side and with line sinks at the front and back. For the purpose of calculation, the cylinder can be imagined to become smaller and smaller so that eventually there are two linear dipoles which have been brought very close together—a quadripole in fact. If we are dealing with a perfect fluid we can write the potential at P due to the doublet S_1N_1 shown in Figure 8 as $(m\cos\theta_1)/r_1$ where m is a quantity proportional to the strength of the doublet. Hence, we can write the potential due to both doublets as $m(\cos\theta_1/r_1 - \cos\theta_2/r_2)$ which, when the doublets approach closely, becomes $m(O_1O_2/r^2)\sin 2\theta$ (since $r_1\cos\theta_1 = r_2\cos\theta_2$). If now we make O_1O_2 extremely small and m proportionately great, so that the product $m \, . \, O_1O_2$ is finite and equal to a, we may write the potential ϕ at P as $\phi = (a/r^2)\sin 2\theta$. The equipotential lines thus form a family of lemniscates. If we require the flow lines, it is only necessary to determine a set of curves which cut the above set orthogonally. Such a set is given by another group of lemniscates

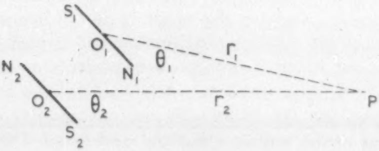


Figure 8.

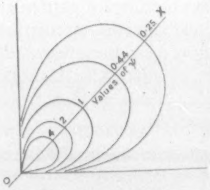


Figure 9. Curves given by $\psi = (a/r^2)\cos 2\theta$.
The zero value of θ is along OX.

$$\psi = (a/r^2)\cos 2\theta \qquad \ldots\ldots(1)$$

where ψ is the flow function.

Laplace's equation in cylindrical coordinates (r, θ) takes the form

$$\frac{\partial^2\psi}{\partial r^2}+\frac{1}{r}\frac{\partial\psi}{\partial r}+\frac{1}{r^2}\frac{\partial^2\psi}{\partial\theta^2}=0$$

and it is clear that it is satisfied by (1). This fact, however, necessarily follows from its method of derivation.

Except near the origin, the lemniscate curves shown in Figure 9 exhibit a very considerable resemblance to the actual stream lines of Figure 4. An interesting imitation, in a perfect fluid, can thus be made of the macroscopic features (for comparatively low Reynolds' numbers) of the circulatory flow around a vibrating cylinder in a real fluid.

On the other hand to deal with the problem at all exactly it is necessary to pass from the equations of a perfect fluid to those of a real viscous fluid, namely the complicated Navier–Stokes equations. This has been done in a rather remarkable paper by Schlichting (1932) to which Professor Andrade first called the author's attention. Schlichting uses a method of successive approximation, with the result that it is found possible to draw diagrams of the circulations that should take place. Schlichting's results are in good general agreement with the experimental investigations of Andrade (1931). The direction of the major circulation is that given by the outer ovals of Figure 5, but, in addition, an indication is given of the inner circulation with a contrary direction of flow, which is dealt with in this paper. Whilst a prediction for a circulation of this kind, based as it is on a difficult calculation, must be regarded as an achievement, it must be pointed out that Schlichting bases his argument on the assumption that the particles trace out linear paths. That this is, in general, not so is shown by Figure 2.

It is interesting to note the close inter-relationship that exists between vibratory and circulatory motion. In this work, attention has been devoted to what occurs around a vibrating reed, but if sound of quite moderate intensity impinges on a diaphragm pierced with a small hole, a vigorous circulatory system somewhat similar in shape to Figure 4 is set up. In the older theoretical calculations (Lamb 1925), relating to the passage of sound through an aperture, no allowance is made for this. Recently (Ingard and Labate 1950) however, the matter has received attention but it would still seem doubtful if the importance of circulations that can occur has been fully allowed for in many measurements and especially in those of the pressure of sound. Indeed, it is felt that the way in which vibratory motion in acoustics is so frequently associated with circulatory motion needs further emphasis.

REFERENCES

ANDRADE, E. N. DA C., 1931, *Proc. Roy. Soc.* A, **134**, 445; 1947, *Nature, Lond.*, **159**, 132.
CARRIÈRE, Z., 1931, *J. Phys. Radium*, **2**, 165; 1940, *Ibid.*, **1**, 68.
INGARD, U., and LABATE, S., 1950, *J. Acoust. Soc. Amer.*, **22**, 211.
LAMB, H., 1925, *Sound*, 2nd Ed. (London: Arnold), p. 249.
LAWSON, R. W., 1947, *Nature, Lond.*, **159**, 168.
SCHLICHTING, H., 1932, *Phys. Z.*, **33**, 327.
SIMPSON, F. W., 1947, *Nature, Lond.*, **160**, 93.
WEST, G. D., 1946, *Nature, Lond.*, **158**, 755; 1947, *Ibid.*, **159**, 167.

A Study of Thermoelectric Effects at the Surfaces of Transistor Materials

By J. W. GRANVILLE and C. A. HOGARTH
Physics Department, The University, Reading

Communicated by R. W. Ditchburn; MS. received 24th November 1950, *and in amended form 26th January* 1951

ABSTRACT. The surfaces of various specimens of Ge and PbS have been explored with a whisker contact, and the character of the conduction mechanism in relatively small regions examined by the polarity of (*a*) the photo-voltaic effect, (*b*) the rectification, and (*c*) the thermoelectric effect. Methods (*a*) and (*b*) give results characteristic of the bulk material as determined by Hall effect measurements, whereas method (*c*) only gives results consistent with (*a*) and (*b*) for uncontaminated cleaved surfaces or for surfaces which have been etched after polishing. For polished surfaces anomalous effects are obtained and point to the existence of a layer on the surface of character different from that of the bulk semiconductor.

---

§ 1. INTRODUCTION

In the course of research on germanium, lead sulphide and other materials which can exist as excess as well as deficit semiconductors, it is often desirable to test relatively small regions for the character of their conduction mechanism. Three properties can serve as the criterion as to whether a given region is predominantly of the excess or deficit type. These are the direction of rectification, the polarity of the thermoelectric effect, and the polarity of the photo-voltaic effect. Theory predicts that when tests are carried out over a clean surface of a semiconducting specimen, there should be complete correlation between the three effects. It was thought important to examine whether this correlation could be observed in practice and to assess, if possible, the relative merits of the three criteria.

A further reason for the investigation concerned the possibility of the presence of surface layers whose characteristics differ from the bulk material. The existence of such layers has been suggested on a number of previous occasions (Bardeen and Brattain 1948, Brattain and Bardeen 1948), for instance, in connection with the Ge transistor and, although it is now known that such layers are not essential to the mechanism of transistor action, further evidence for or against their existence was considered desirable.

With these objectives in view, a number of Ge and PbS surfaces were examined in the light of the above three criteria. The results have shown that complete correlation between these effects is not always obtained on the specimens investigated, particularly in the case of surfaces which have been polished before testing.

§ 2. EXPERIMENTAL METHOD

For the investigation of the various surfaces the following simple arrangement was used.

The specimen, which carried a low-resistance base electrode, was mounted on a slide with which were incorporated two perpendicular micrometer movements in a horizontal plane. A whisker contact was mounted on a fixed pivot, resting in jewelled bearings. The total thrust on the contact points could be regulated by means of a smoothly movable weight and could thus be adjusted from nearly zero to 9 gm. weight. The diode characteristic of the contact could be displayed on an oscilloscope using a conventional circuit. To test the polarity of the thermoelectric effect, the whisker could be electrically heated by means of a heater coil incorporated in the mounting head (Figure 1) and the test circuit completed through a galvanometer. A temperature difference was thereby established between the whisker and the base electrode and the polarity of the resulting thermo-current was noted from the direction of the galvanometer deflection. The *magnitude* of this deflection was not considered as having any important significance in view of the ill-defined contact conditions and the dependence of thermo-currents on the contact resistance. It was expected that the *polarity* of the thermo-current would be the same as that of the thermo-E.M.F., and observations with a potentiometer confirmed that this was so. Thus, in view of the large number of observations to be made, and since no experimental reliability was lost by doing so, a galvanometer method was employed for the observation of thermoelectric effects. To test the photo-voltaic effect, light was focused on the contact point by means of a simple projection system and the resulting photo-current measured on a galvanometer. Suitable precautions were taken to exclude heat radiation from the source. The whisker was prepared from 0·005 in. tungsten wire, pointed by electrolytic etching to a radius of curvature of less than 0·0005 in. The experiments involving heat treatment were performed in a silica tube connected to a glass vacuum system which gave a pressure of the order of 10^{-5} mm. Hg. The germanium used was obtained from various sources and contained different impurities but the contact properties described here were found to be representative of all specimens.

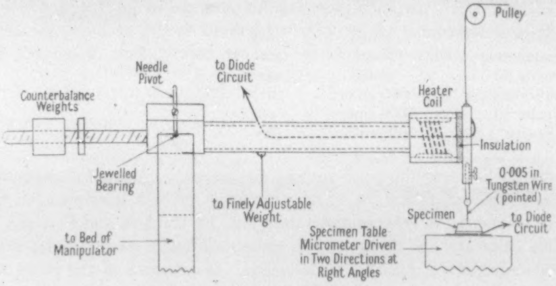


Figure 1. Mounting head and pivot arm for contact point.

Some of the Ge surfaces used in these experiments were cleavage planes of single crystals or cleaved surfaces of polycrystalline specimens. Polished surfaces were obtained by a variety of methods as given in the Table. The results described below were found to be independent of the method of polishing and thus contamination of the Ge surface by the abrasive material cannot be considered responsible for the observed anomalies in the thermoelectric effect. The PbS specimens used were taken from photo-voltaic cells of German origin and were of the type previously used for experiments on the PbS transistor (Gebbie, Banbury and Hogarth 1950).

Results of Thermoelectric Tests after various Methods of Polishing

Grinding abrasive	Polishing abrasive	Conditions of polishing	Results of thermoelectric tests on surface
Aloxite 50	—	water	Some p-type indications, but mostly n-type
Carborundum 700	—	water	,,
Carborundum 700	rouge	dry	All p-type at low contact temperatures
Aloxite 50	rouge	dry	,,
Carborundum 700	rouge	water	,,
Aloxite 50	rouge	water	,,
Carborundum 700	0000 emery	dry	,,
Aloxite 50	0000 emery	dry	,,
1 Emery	0000 emery	dry	,,

The temperature at the point contact was estimated in the following manner. The thermoelectric power of a representative etched specimen of Ge was accurately measured in the apparatus described by Henisch and François (1950). From this value and the thermo-E.M.F. measured with a heated point contact as described above, it was possible to assess the temperature of the point and this was calibrated against the current flowing in the heater circuit. Thus in subsequent experiments the approximate temperature of the point was inferred from the heater current. The method was checked by attaching a fine thermocouple to the whisker very close to the junction.

§ 3. EXPERIMENTAL OBSERVATIONS

The results obtained with the above arrangement may be briefly summarized as follows:

(*a*) All Ge surfaces, whether polished or cleaved, when tested for direction of rectification showed that the condition for easy current flow was obtained when the Ge was negative with respect to the whisker, as expected for an n-type semiconductor. All PbS surfaces, whether polished or cleaved, showed the opposite behaviour, as expected for a p-type semiconductor.

(*b*) On all surfaces of Ge the photo-E.M.F. was such that the whisker was positive with respect to the base electrode as expected for an n-type semiconductor. The converse behaviour was found with the PbS specimens. It was found that for both Ge and PbS the average photo-E.M.F. obtained when a cleaved surface was illuminated was about six times the value obtained when a polished surface on the same crystal received the same illumination.

(*c*) Thermoelectric tests produced certain anomalous results. Cleaved surfaces of Ge showed thermo-E.M.F's in the sense expected for an excess semiconductor for all contacts: cleavage planes of PbS indicated deficit conduction. This was to be expected and showed complete correlation with (*a*) and (*b*). On polished surfaces of both Ge and PbS, thermo-E.M.F's of either polarity could be obtained in different regions. Figure 2 shows a 'map' of a surface of this kind for constant heating conditions of the contact probe. These results could not be reproduced point-for-point, probably because the surface properties of the crystal change significantly over distances smaller than could be set by the manipulator. Regional configurations were, however, found to be reproducible.

(*d*) The configurations obtained depended on the temperature of the contact probe. Figure 3 shows the thermoelectric current for a particular contact spot on a polished Ge surface, plotted against the temperature of the contact. It will be seen that such a spot would be recorded as p-type for low temperatures and as n-type for high temperatures. This effect is completely reversible with temperature. Similar results were obtained when thermo-E.M.F's were produced between two contact points at different temperatures placed on the same polished surface. Figure 4 shows the results of repeated thermoelectric mapping of a given polished surface at various contact temperatures. The fraction of the total area recorded as p-type tended to unity as the temperature of the point contact decreased, and tended to zero as the temperature increased. The opposite behaviour is found on polished surfaces of PbS as shown in Figure 3. For low probe temperatures a particular contact spot would be recorded as n-type, but as p-type for high temperatures. This behaviour also was reversible with contact temperature. Figure 4 also shows the effects of repeated thermoelectric mapping of a polished PbS surface. The variation of thermo-E.M.F. with temperature was confirmed by measuring with a potentiometer instead of a galvanometer, when results similar to those in Figures 3 and 4 were obtained.

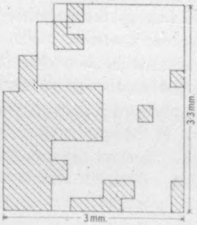


Figure 2. The nature of the surface conductivity of a germanium sample, as determined by the thermoelectric technique. Shaded areas: p-type; unshaded areas: n-type.

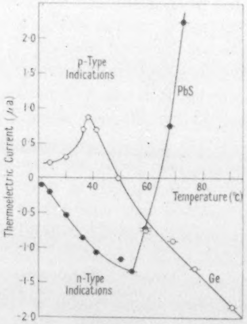


Figure 3. Thermoelectric current as a function of contact temperature for polished surfaces of Ge and PbS. (Base connection at about 20° c.)

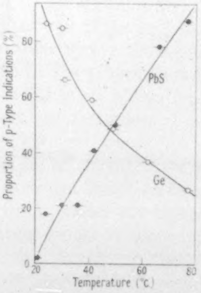


Figure 4. Analysis of the results obtained by repeated thermoelectric mapping of a Ge and a PbS surface for different contact temperatures. (Base contact at 20° c.)

(*e*) Experiments using various materials for the contact probe yielded similar results, i.e. the probe material was not directly responsible for the anomalous thermoelectric effect.

(*f*) When a polished Ge surface of *mixed* character was etched (using a mixture of HF, HNO_3, and water, containing a small quantity of $Cu(NO_3)_2$) all regions which were initially p-type as shown by the thermoelectric tests changed to n-type. The character of the etched surface was then uniformly n-type.

Conversely, when a polished surface of PbS was etched (using concentrated HCl) all regions which were initially n-type changed to p-type. The character of the etched surface was then uniformly p-type.

(*g*) Polished Ge specimens were vacuum baked at various known temperatures for $1\frac{1}{2}$ hours. After each heat treatment a given surface area was mapped by the thermoelectric test at a constant probe temperature (about 4° above that of the base). Eighty-one spots for each baking temperature were tested to give a measure of the proportion of p-type indications. The results for six polished Ge surfaces are shown in Figure 5. It is seen that the Ge surfaces became uniformly n-type after treatment at 900° c., that is, a little below the melting point. It is of interest to note the double maximum recorded on five of the six curves and the minimum at about 500° c. If the specimen is heated directly to 900° c. from room temperature, the 'mixed' surface becomes uniformly n-type in less than an hour.

A cleaved surface on each specimen was tested throughout the vacuum baking experiments and the thermo-E.M.F. was always characteristic of an n-type material for all temperatures. Thus the changes observed may be inferred to be characteristic of the polished surfaces and not of the bulk material.

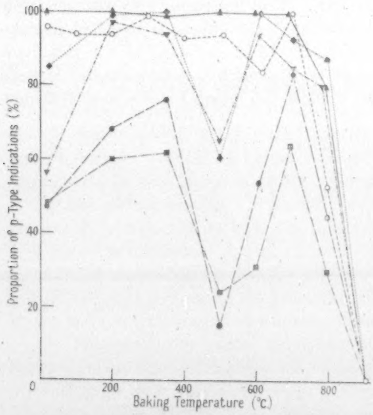


Figure 5. The proportion of p-type indications for Ge surfaces, after baking at various temperatures.

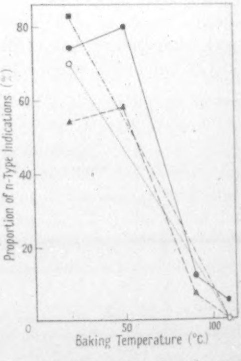


Figure 6. The proportion of n-type indications for PbS surfaces, after baking at various temperatures.

Similar experiments were carried out on PbS specimens (Figure 6). Here the polished surface became uniformly p-type on vacuum-baking at 110° c.

(*h*) When the heat-treated specimens were left in air there was no tendency for the polished surfaces to revert to their original mixed character. The surfaces remained n-type for Ge and p-type for PbS for test periods of at least two months.

(*j*) Electron diffraction patterns were obtained by reflection for the two types of surfaces. It was found that distinct patterns of Kikuchi lines and spots were only obtained on plane cleavage surfaces. On polished surfaces they were not observed. Instead the very broad, diffuse, haloes characteristic of an amorphous substance or polycrystalline material of small grain size were obtained.

§4. DISCUSSION OF RESULTS

The above results for Ge and PbS surfaces show that the predicted complete correlation between the direction of rectification, the polarity of the thermo-E.M.F. and the polarity of the photo-voltaic effect is only observed on uncontaminated cleaved surfaces and on surfaces which have been etched after polishing. No correlation is found on polished surfaces. This suggests that polished surfaces are covered with a layer of material of character different from the bulk material and that this layer can be removed by etching or by a suitable heat treatment. Results of the type shown on Figure 2 suggest that this postulated surface layer varies from place to place in thickness or in character or both. It is concluded that except for this layer the Ge specimens used are excess semiconductors and that, conversely, the PbS specimens must be regarded as deficit materials. This view is confirmed by the sign of the Hall effect. On this basis, the photo-voltaic and rectification tests may be considered reliable in every case whereas the results of thermoelectric surface tests must be examined with caution.

Two important problems arise in connection with the above hypothesis, namely, what is the nature of the postulated surface film, and how can it account for the reversal of thermoelectric power and for its complicated temperature dependence. The second of these is particularly difficult, since, on general grounds, surface layers—unless they themselves provide a source of E.M.F.—should have no effect on the *polarity* of the thermo-E.M.F., unless the conditions are such that a considerable temperature difference can exist across them. In the present case, this condition could not be satisfied unless the films had a thermal conductivity several orders of magnitude lower than that of the bulk material.

A first hypothesis ascribed the phenomena in some way to the adsorption of gas and its dynamic equilibrium with the atmosphere. The experiments described under (*g*), (*h*) and (*j*) above failed to confirm any mechanism of this type.

It is probable that the act of polishing causes the formation of a layer of amorphous material on the crystalline Ge. Additional impurities may be introduced and the layer may represent a quasi-liquid arrangement of atoms (Beilby layer) as is thought to exist on certain polished metal surfaces. It is reasonable to suppose that prolonged heating would cause the crystal lattice to grow at the expense of this layer, thus accounting for the results summarized under (*g*).

The electron diffraction tests referred to above are consistent with the idea of an amorphous layer at the surface of a polished specimen (see Thomson and Cochrane 1939). No evidence of surface contamination or modification is evident on cleaved surfaces, whereas the very diffuse haloes produced when polished surfaces are examined indicate the presence of an amorphous surface or one which consists of a conglomerate of very small grains.

In order to elucidate the nature of this surface layer, a subsidiary experiment was carried out as follows. A thin layer of Ge was vacuum-evaporated on to a cold cleavage plane of Ge, under conditions of evaporation which had always previously yielded amorphous layers of Ge when vacuum-evaporated on to unheated glass substrates. The thickness of the layer was estimated optically as approximately 10^{-5} cm. The thermo-E.M.F. over this surface was then investigated, and it was found that the previous strong n-type indications were no longer observed. At low temperatures of the contact probe the thermoelectric currents were too small to be conveniently observed, while at high temperatures the tests yielded feeble n-type indications. The presence of the evaporated layer also suppressed the photo-voltaic effect, whereas the rectification characteristics were not substantially different from those observed in absence of the films, although the values of peak back-voltages had decreased. It cannot be supposed that a layer of this type, deposited by vacuum evaporation, will have exactly the same structure as one produced by mechanical polishing, but the experiment does demonstrate that the presence of thin surface layers whose structure differs from that of the bulk material can have a profound effect on electrical measurements. The amorphous films are expected to have a lower thermal conductivity, though from quantitative considerations it seems improbable that this difference is large enough to account by itself for the observed effects.

With a view to determining at which junction the anomalous thermoelectric effects occur, a subsidiary experiment was carried out using two point contacts. A cold point was placed on a polished Ge surface which had previously shown anomalous (p-type) thermo-E.M.F's, and a heated contact point was placed on an etched surface of the crystal at a considerable distance away. The maximum temperature gradient was thus expected to be situated at the contact on the etched surface. The E.M.F. was such that the hot junction was of positive polarity, that is n-type thermoelectric indications were obtained. The experiment thus showed that the anomalous effects are only observed when the greatest fall of temperature occurs at the junction on the polished part of the surface.

Thus, although these results are not completely understood, they provide additional experimental data on the properties of metal–semiconductor contacts. They show also that the results of tests made with the hot-point technique, which has been used by various workers for the investigation of the nature of conductivity in semiconducting materials, must be interpreted with caution.

ACKNOWLEDGMENTS

The authors would like to express their thanks to Professor R. W. Ditchburn for the provision of research facilities and for his encouragement, and to Dr. H. K. Henisch for his co-operation and many stimulating discussions.

REFERENCES

BARDEEN, J., and BRATTAIN, W. H., 1948, *Phys. Rev.*, **74**, 230.

BRATTAIN, W. H., and BARDEEN, J., 1948, *Phys. Rev.*, **74**, 231.

GEBBIE, H. A., BANBURY, P. C., and HOGARTH, C. A., 1950, *Proc. Phys. Soc.* B, **63**, 371.

HENISCH, H. K., and FRANÇOIS, M., 1950, *Proceedings of the International Conference on Properties of Semiconducting Materials* (London: Butterworth's Scientific Publications).

THOMSON, G. P., and COCHRANE, W., 1939, *Theory and Practice of Electron Diffraction* (London: Macmillan).

The Deformation and Ageing of Mild Steel

By W. SYLWESTROWICZ[*] and E. O. HALL
Cavendish Laboratory, Cambridge

Communicated by W. L. Bragg; MS. received 18th January 1951

ABSTRACT. Mild steel deforms by a series of bands of plastic deformation called Lüders bands. In heavy tensile specimen a complexs series of bands may arise, but in thin wire specimens it is shown that only single bands are formed. Measurements have been made of the propagation stress and Lüders strain for these single bands, and the effects which occur on ageing deformed specimens have been studied.

A critical examination of current theories of the upper yield point—the stress at which bands first appear—is also undertaken.

§ 1. INTRODUCTION

MILD steel does not yield uniformly but by the formation and subsequent broadening of bands of plastic deformation, which have been variously called flow lines, strain figures, or stretcher strains, though more generally Lüders lines, after their discoverer, W. Lüders (1860).

A diagrammatic stress–strain curve for the material is reproduced in Figure 1. From A to B the deformation is elastic, but at B, called the upper yield point, the stress suddenly falls, with the simultaneous appearance across the specimen of the first Lüders band. Outside the band the material is undeformed. The bands then spread over the specimen under the lower yield stress (the ordinate of the line CD) until, at the point D, the entire specimen is covered. The strain at D will be called the Lüders strain. At this point the stress rises sharply, and from then on the deformation is homogeneous.

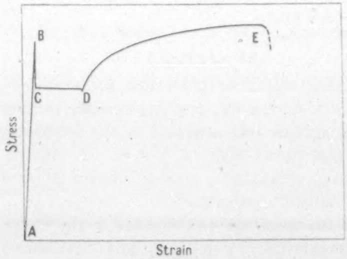


Figure 1. Diagrammatic stress–strain curve of mild steel.

With the normal type of heavy tensile specimen, a complex series of bands may arise. Under these conditions observations on the lower yield point are very difficult, because the stress fluctuates sharply as each new series of bands is formed. The Lüders deformation entails a shear, so that the deformed and undeformed sections of the test-piece are no longer co-planar. Consequently, the rigidity of the test-piece, and the constraints imposed by the grips, play a large part in determining the form of the bands, and it might be expected that the lighter the load, and the thinner the specimen, the more chance a single band has of propagating uniformly from one end of the specimen to the other. This was found to be the case in a preliminary investigation carried out at the Cavendish Laboratory by Dr. E. Orowan and Mr. W. Chitruk. They found that in thin wire specimens, only one or at most two Lüders bands were formed, one from each grip, and spread under virtually constant stress over the specimen. Thus, accurate measurements of the propagation stress and Lüders strain are possible.

---

[*] Now at the Mechanical Engineering Research Organization, National Physical Laboratory, Teddington, Middlesex.

§ 2. MATERIAL

The wires used in these experiments were 1 or ½ mm. in diameter, and were kindly supplied by Messrs. Rylands Bros. Ltd., Warrington. The analysis of the three steels was as shown in Table 1. From the coils of hard-drawn wire, lengths of about 25 cm. were cut, straightened, annealed *in vacuo* until the necessary grain size was obtained, and then furnace-cooled. From the middle of the 25 cm. specimen a 14 cm. length was cut. This made a 10 cm. gauge length with 2 cm. in each grip.

Table 1. Analysis of Iron Samples

Material	Impurities present (%)						
	C	S	P	Mn	Ni	Cu	Si
Armco	0·020	0·025	0·009	0·035	0·036	0·078	—
Siemens-Martin (PXQ quality)	0·060	0·036	0·022	0·25	—	—	0·014
Basic Bessemer (Thomas quality)	0·045	0·047	0·04	0·33	—	—	Trace

§ 3. APPARATUS

The testing machine was a simple frame set vertically, consisting of two uprights, between which moved the straining spindle, guided by four adjustable conical rollers. The spindle was attached to the lower end of the specimen. The motive power was provided by a ¼ H.P. D.C. motor, and with suitable electrical and gearbox connections, a wide choice of crosshead speeds was available—from 0·02 mm/hr. upwards.

The upper end of the specimen was attached to the stress measuring system. In the earlier experiments (W.S.) the load was determined by a set of eight electric alstrain gauges, in later work (E.O.H.) by the deflection of a spring-steel plate. The extension of the specimen could be calculated from the readings of a revolution counter attached to the gear train.

For normal tensile tests, or for studies on the propagation of the bands, the wire specimens were clamped between small grips. If investigations were to be made on the value of the upper yield point, then the stress concentrations around these grips had to be avoided, either by depositing heads on the wires electrolytically, or by a system of solder cups. These cups were of brass, containing a ⅛ in. central hole, filled with 'Matrix' alloy, a tough brittle alloy with the composition of Woods metal, but containing antimony instead of cadmium. Its melting point is 104° C. The wires were frozen into these cups, and tested in the normal manner.

§ 4. EXPERIMENTAL RESULTS

(i) *The Upper Yield Point*

The upper yield point is the stress at which the first Lüders band appears. In this study of the factors which affect its value, attention will be directed on two: the result of stress concentrations, and the grain size of the specimen.

Lüders lines always start from the grip or from the fillets of a tensile test-piece. These inhomogeneities may be removed in many ways. In shaped specimens the fillets may be made with very large radii of curvature. In this way, Haigh (reported by Edwards *et al.* (1943)) obtained an upper yield point 18% above the ultimate tensile strength. Such shaped specimens, however, are difficult to use for stress–strain measurements, as the gauge length is continuously changing while the stress increases.

In wire specimens the stress concentrations at the grips can be avoided by depositing heads on the wires electrolytically as mentioned in the previous section. In this way, one of us (W.S.) was able to obtain an upper yield point almost 100% above the lower yield stress when a Lüders band formed in the middle of the specimen. This yield stress would undoubtedly be well above the ultimate strength for material of this grain size. For more routine investigation, however, the system of solder cups was used. With them a wide range of material of varying grain sizes was examined, in an attempt to find the dependence of the upper yield point on the grain size of the matrix.

These cups do not give such gently tapering specimens as the electrolytic heads, but the specimens were sufficiently regular to enable any trends in the yield point to be observed. Obviously, when making these measurements, the *highest* value obtained must be taken, for this implies the maximum freedom from inhomogeneities.

The results are displayed in Table 2. It will be seen that the upper as well as the lower yield points are strongly sensitive to grain size, and that the difference between the two drops to a very small value in the larger grained specimens.

Table 2. Upper and Lower Yield Point Values

Armco—Grain size (No/mm.)	133	61	46	18	
σ_{UYP} (kg/mm^2)	33·8	24·2	19·2	13·7	
σ_{LYP} (kg/mm^2)	26·5	20·4	16·9	12·9	
Thomas—Grain size (No/mm.)	135	99	49	27	
σ_{UYP} (kg/mm^2)	42·8	26·5	23·0	20·4	
σ_{LYP} (kg/mm^2)	27·4	24·0	19·7	19·2	
PXQ—Grain size (No/mm.)	61	60	53	31	24
σ_{UYP} (kg/mm^2)	30·1	27·9	26·3	19·4	14·2
σ_{LYP} (kg/mm^2)	20·7	19·7	17·3	15·4	13·3

These tests were all carried out at the slow strain rate of 10^{-4}/min., so that down to this level there is no indication that extremely slow loading will cause the yield point effect to disappear. This is contrary to the results of Elam (1938) on Armco iron.

(ii) *The Lower Yield Point*

Once the Lüders band has formed, the load falls abruptly to the lower yield stress, and further deformation takes place by the broadening of the band. The local stress concentrations at the band edge assist the applied load to reach the yield stress of the undeformed material, and in this way the band spreads over the specimen. As each new lamella yields, it flows under the applied stress by creep. Since the initial extension in a creep test takes place extremely rapidly, practically the whole of the deformation will occur in the lamella next to the edge of the band.

By measuring the diameter of the wire before and after the extension, it can be shown that in the Lüders band the cross section has become elliptical. The minor axis of the ellipse is, in specimens with about 60 grains/mm., some 2 to 2·5% and the major axis 1 to 1·5% smaller than the original diameter; thus, the total strain in the Lüders band is about 2·5 to 3·5%, depending on the rate of extension.

The strain rate has a profound effect on both the lower yield stress and the Lüders strain. This is shown in Figure 2. The lower yield stress rises rapidly as the deformation rate increases, and the Lüders strain rises in sympathy.

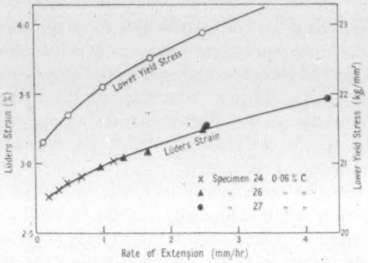


Figure 2. Variation of propagation stress and Lüders strain with strain rate.

These points have already been noted by Elam (1938) and others, but the measurements made here have one important difference: they could all be made on a single band propagating in one specimen. It is not even necessary to wait to the end of the yield point extension to determine the Lüders strain ϵ: this can be found if the velocity of the band edge v' and the velocity of the straining head V are known. For during time dt, a length $v'\,dt$ of the wire extends to a length $v'\,dt(1+\epsilon)$. But in this time the straining head moves a distance $V\,dt$. Hence $v'\,dt(1+\epsilon)-v'\,dt=V\,dt$, so that $\epsilon=V/v'$. It is possible to see the boundary of the band by careful adjustment of the illumination, and measurements of its velocity are conveniently made with a travelling microscope. The results are as shown in Figure 2.

(iii) *Strain Ageing Phenomena*

With the appearance of the first Lüders band, the upper yield point as observed on the testing machine disappears, for if the specimen is unloaded and immediately reloaded, yielding continues at the lower yield stress. However, if the specimen is left for some time at room temperature, the deformed material in the band hardens by strain ageing, and an additional stress must be applied before the band front spreads further. This gives rise to a small upper yield 'pip' in the stress–strain curve after ageing, as shown in the inset to Figure 3.

The growth of this 'pip' may be accelerated by ageing at slightly elevated temperatures, and the height of the 'pip' increases the longer the ageing time. This variation is shown schematically in Figure 3.

The results are for a specimen of PXQ quality iron with 132 grains/mm. After the band has restarted, the stress falls to the original lower yield stress level. By repeating the ageing process on the same specimen, scatter in the results is cut to a minimum.

After more severe ageing than indicated in Figure 3, the whole stress–strain curve changes. It appears that the front of the band is now so effectively locked that it cannot move at all. Instead, deformation occurs by a new Lüders band which is nucleated again at the grip, and spreads through the *old* strain-aged band. This new band has already been named the 'secondary' band, and the original band the 'primary' one, in a brief report on these ageing phenomena (Hall 1950).

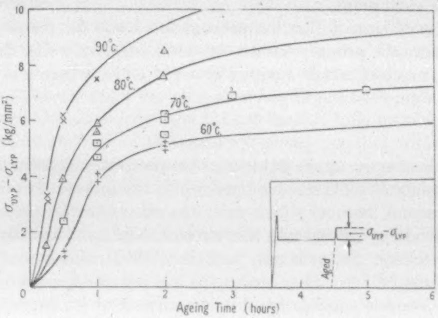


Figure 3. Growth of upper yield point on ageing.

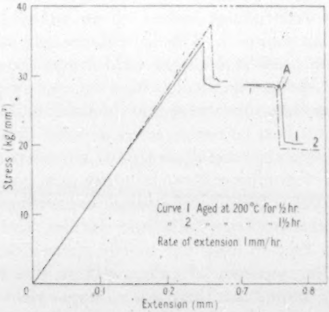


Figure 4. Secondary bands in aged material.

Two typical curves are shown in Figure 4. The material was the same as for Figure 3, but with a grain size of some 60/mm. In each case the combination of the tensile stress and grip pressure nucleates the secondary band at the grip from which the primary band began. This band propagates through the primary at a much higher stress level. When the front of the secondary reaches the boundary of the primary at the points A, the propagation stress drops to the normal value again. It is surprising that the material prefers to deform in the strain-aged area, rather than to spread the primary band further into the still undeformed part of the specimen. The change-over from primary to secondary band propagation depends on ageing temperature, time of ageing, and also possibly on grain size. For instance, with the finer-grained material of Figure 3, ageing at 90° c. for more than one hour was always sufficient to start a secondary band, but in the coarser-grained material higher temperatures were necessary. With the lower ageing temperatures a marked amount of plastic deformation could occur before the yield point.

It is seen from Figure 4 that the propagation stress for the secondary band may be 40% above the primary band. It seems surprising that the yield stress may be raised so much after, say, a mere 3% strain, and any theory of strain ageing will have to account for such large changes in the yield stress from such small deformations.

§ 5. DISCUSSION

The drop in load at the upper yield stress has been the subject of a considerable amount of speculation. There have been, up to the present, two main theories for this phenomenon, both of which refer the effects to the presence of small quantities of carbon or nitrogen in a matrix of α-iron. The first is the older grain-boundary theory, due primarily to Dalby (1913), and extended by Nadai (1924) and Kuroda (1938). Here the grains are imagined surrounded by some form of harder material, which has become identified in the minds of many workers with cementite. This boundary film is supposed to withstand the deformation of the grains, until the stress reaches the upper yield point; deformation can then spread across the specimen, and the first Lüders band is formed. The film reforms during the ageing treatment by the diffusion of new carbon. The second theory is due to Cottrell (1948), and developed more fully by Cottrell and Bilby (1949). In this case the dislocations normally present in unstrained material are looked on as locked by the carbon and nitrogen atoms present in solution. The dislocations are pulled away from their 'atmospheres' of carbon atoms at the upper yield stress; the atmospheres reform during any subsequent ageing treatment, with a corresponding hardening.

For Cottrell's theory to be valid, three experimental results must be expected: (i) the upper yield point must be destroyed by any free dislocations introduced into the material, (ii) the upper yield point must be independent of grain size, and (iii) single crystals of iron must show an upper yield point.

The question of an upper yield point in single crystals has given rise to much controversy. Edwards and Pfeil (1925) and Holden and Hollomon (1949) failed to find any such drop in load, but shortly afterwards, Schwartzbart and Low (1949) reported the opposite. Further work in this country by Cottrell and Churchman (1949) showed that there was no upper yield point in unstrained single iron crystals, but a small yield point could be introduced after straining and ageing. Without discussing these experiments in detail, it may be said that the upper yield point in single crystals is either extremely small or non-existent, and by no means comparable with the drop in load in fine-grained material.

Secondly, the results of the previous section have shown that the upper yield point is strongly sensitive to the grain size of the specimen.

Finally, certain experiments have been made with cuts and scratches, with the object of introducing free dislocations into the specimens. The specimens were tested immediately after treatment, so that the carbon atmospheres had no time to form. Light file marks appeared to have little or no effect on the upper yield point, although the lightest scratch must contain many thousands of free dislocations. With heavier cuts and nicks, the yield point was definitely lowered, so that the Lüders bands began from this strained area, but a distinct drop in load was still present. In fact, every specimen was clamped between grips, where considerable plastic deformation occurs, but small upper yield points were still observed.

The fact that a Lüders band is formed in an area of stress concentration seems to indicate that a Lüders band spreads from a small region of recent plastic deformation in the specimen, the enlargement of the plastic area taking place at the upper yield stress. Once the area has begun to spread, the stress falls to the lower yield level, and the first Lüders band is fully formed.

If the number and size of these plastic 'nuclei', as they will be called, are at all large, then the stress–strain curve should show a marked deviation from linearity in the elastic region. However, no such deviation was found in these experiments. By scratching gauge marks on a specimen, it is easy to show that no macroscopic deformation occurs outside the Lüders band. Areas of plastic deformation too small to influence these macroscopic measurements could of course be detected by back-reflection X-ray pictures. Cowley and Paterson (1947) did in fact find a few such spots, but their results have not been confirmed in recent work by Finch (1950). No X-ray evidence has been obtained at this laboratory for deformation outside the clearly defined Lüders bands on our specimens, so that it appears the number of these small plastic nuclei must be very small indeed.

Before summarizing the results of this section, it must be mentioned that support has been obtained for Cottrell's theory by the thermal hardening and the appearance of yield points in single crystals of zinc containing dissolved nitrogen (Wain and Cottrell 1950). Iron single crystals have a small lower yield point extension (Cottrell and Churchman 1949), but none of these effects is on the same scale as any observed in fine-grained mild steel.

Thus, the general conclusion of this section must be that Cottrell's theory of the yield point is unlikely, and that some form of grain-boundary theory is needed to explain the results. This work gives no indication of the nature of any grain boundary film, but merely shows the important influence of the number of grain boundaries on the properties of mild steel. It has also been suggested here that a nucleation hypothesis will cover the existence of the drop in load at the upper yield point.

Further experimental results and a full discussion of this theory will take place in a later paper.

ACKNOWLEDGMENTS

This work was begun (by W.S.) for the British Iron and Steel Research Association. It was continued by the second author under scholarships from the Lord Rutherford Memorial Fellowship, and from the Royal Commission for the Exhibition of 1851. Grateful thanks are tendered to all these bodies for their financial assistance.

Both authors also wish to acknowledge the continued interest and encouragement of Dr. E. Orowan during the development of this work.

REFERENCES

Cottrell, A. H., 1948, *The Strength of Solids* (London : Physical Society), p. 30.

Cottrell, A. H., and Bilby, B. A., 1949, *Proc. Phys. Soc.* A, **62,** 49.

Cottrell, A. H., and Churchman, A. T., 1949, *J. Metals* (*Metals Trans.*), **185,** 877.

Cowley, J. M., and Paterson, M. S., 1947, *Nature, Lond.*, **159,** 846.

Dalby, W. E., 1913, *Proc. Roy. Soc.* A, **88,** 281.

Edwards, C. A., and Pfeil, L. B., 1925, *J. Iron Steel Inst.*, **112,** 79.

Edwards, C. A., Phillips, D. L., and Liu, Y. H., 1943, *J. Iron Steel Inst.*, **147,** 153.

Elam, C. F., 1938, *Proc. Roy. Soc.* A, **165,** 568.

Finch, L. G., 1950, *Nature, Lond.*, **166,** 508.

Hall, E. O., 1950, *Proc. Phys. Soc.* B, **63,** 724.

Holden, A. H., and Hollomon, J. H., 1949, *J. Metals* (*Metals Trans.*), **185,** 179.

Kuroda, M., 1938, *Sci. Pap. Inst. Phys. Chem. Res., Tokyo*, **34,** 1528.

Lüders, W., 1860, *Dinglers polytech. J.*, **155,** 18.

Nadai, A., 1924, *Z. Tech. Phys.*, **5,** 371.

Schwartzbart, H., and Low, J. R., 1949, *J. Metals* (*Metals Trans.*), **185,** 637.

Wain, H. L., and Cottrell, A. H., 1950, *Proc. Phys. Soc.* B, **63,** 339.

On the Radio-Frequency Requirements of High Energy Electron Synchrotrons

By T. R. KAISER*

The Physical Laboratories, University of Manchester

Communicated by P. M. S. Blackett; *MS. received 23rd November* 1950, *and in amended form 8th February* 1951

ABSTRACT. A theory of the transition between betatron and synchrotron acceleration is specifically applied to high energy electron synchrotrons. The results are used to calculate the R.F. voltage required for two such machines at present being constructed. It is shown that with suitable choice of its initial rate of rise, the R.F. voltage need rise only to a peak value which is not greatly in excess of the voltage per revolution required to maintain an electron on the synchrotron equilibrium orbit.

§ 1. INTRODUCTION

Wilkins (1950) has considered in detail the problems associated with betatron-starting in high energy electron synchrotrons (peak energies above 100 Mev.). Amongst other factors, he has calculated the radio-frequency accelerating voltage needed for successful operation of a specific machine (the Glasgow University 375 Mev. accelerator). Wilkins followed, in this, the theory of Goward (1949) which assumes the R.F. voltage to be instantaneously established.

The purpose of this communication is to apply the theory of electron capture of Kaiser (1950), which takes into account the finite rate of rise of the R.F. voltage, to the energy range considered by Wilkins. It is found that this will result in a reduction in the voltage calculated to be necessary for successful electron capture, over that derived by Wilkins, with consequent saving in R.F. power, and also in a relaxation in the restriction on the electron energy at transition, set by losses due to electron collisions with the boundaries of the orbit space, thus allowing greater flexibility in betatron flux bar design.

---

* I.C.I. Research Fellow.

The notation of the previous paper (Kaiser 1950, to be referred to as I *) is rather cumbersome because of the multiplicity of symbols employed and so will be slightly modified. All symbols are defined to simplify comparison with the previous results.

r (cm.) = equilibrium orbit radius, n = magnetic field index (see I), E (ev.) = total electron energy (including rest energy), ϕ_s = equilibrium phase (see I), β = ratio of electron velocity to that of light, ω (sec^{-1}) = angular velocity of the equilibrium electrons = angular radio frequency, v (volt) = that part of the total accelerating voltage per revolution required to maintain a particle on the equilibrium orbit, not supplied by betatron induction, V (volt) = peak R.F. voltage, $\sin\phi_s = v/V$, and $\Phi = \cos\phi_s - (\frac{1}{2}\pi - \phi_s)\sin\phi_s$.

The subscripts i and T refer to the values of the parameters at the instants of injection into the betatron phase and of transition to synchrotron acceleration respectively.

The important qualitative results obtained in I were, (*a*) that by allowing the peak R.F. voltage V to increase sufficiently slowly to a value which we will call V_0, effectively all particles will be captured which have betatron instantaneous orbits within a region of radial extent $1/\sqrt{2}$ times the width of the synchrotron stable region corresponding to the voltage V_0, and (*b*) that in the case of the high energy machines here considered it is possible to capture, and retain during betatron saturation, all particles from the betatron beam by allowing V to rise to V_m which is appropriately greater than V_0.

§ 2. THEORY

Suppose that when $V = V_0$ the stable state is just established, that is, all electrons are captured and are executing phase oscillations with amplitudes from zero to 180° (assuming that the betatron flux condition is satisfied at the instant of transition). As V increases further, the phase oscillations will be damped approximately as $V^{-1/4}$. This bunching of the electrons has been considered by de Packh and Birnbaum (1948) as a means of minimizing the peak amplitude of the radial oscillations.

Using the theory of I we can show that this damping will be adequate to prevent loss of particles during the subsequent betatron saturation if V increases to a value V_m given by

$$\frac{\Phi_1^2}{\tan\phi_{s1}} \simeq \frac{E_T V_0}{E_1 v_1} = C \text{ (say)}, \qquad \ldots\ldots(1)\dagger$$

where we assume that v increases from a value near zero ($v/V_0 \ll 1$) to a maximum value v_1, corresponding to $E = E_1$, $\phi_s = \phi_{s1}$, $\Phi = \Phi_1$, during the period of betatron saturation ($\sin\phi_{s1} = v_1/V_m$).

The effect of phase damping with increasing magnetic guide field is represented by the factor E_T/E_1. The true factor will depend to some extent on the manner in which v increases and is not easily calculable. It will certainly lie between unity and E_T/E_1 and, as the examples will show, within wide limits the value assumed will not greatly influence the results.

---

* *Note.* The title of Figure 7 of I should read, '. . . assuming v/V to increase linearly from 0 to 0·5 . . .'. Lines 19 and 29 on page 61 should read, '. . . we see that $[\dot{\phi}_m/(2kV)^{1/2}]E_s^{+1/4}$ remains constant . . .', and '. . . increases linearly from 0 to 0·5 . . .', respectively. Equation (29) should read $\dot{\phi}_m = (E_{sp}/E_s)^{1/4}(4kV\Phi_p)^{1/2}$.

† In I the phase oscillation damping during the period of betatron saturation was assumed to be as $E^{-1/4}$. If $\cos\phi_s$ decreases substantially from unity during this period, the factor should be $(E\cos\phi_s)^{-1/4}$. Making this change, (1) is obtained from equation (31) of I.

The solution of equation (1) for V_m/v_1 is given in Figure 1, where (V_m/v_1-1) is plotted against C.

Radiation losses from the circulating electrons are neglected. They will not be significant at energies of the order of E_1 and phase damping will prevent loss of electrons from this source provided that V_m is sufficiently greater than the maximum value attained by v when radiation losses are included.

In order that the transition efficiency should approach 100% the following conditions are found from equation (33) and from equations (7) and (24) of I, respectively:

$$2t_0[\omega^2 V_0/\pi E_T(1-n)]^{1/2} \gg 1, \qquad \ldots\ldots(2)$$

$$V_0 > \pi(1-n)E_T(\rho/r)^2 = V_0'. \qquad \ldots\ldots(3)$$

The electrons are assumed to be extreme relativistic, ρ is the half spread of the betatron beam immediately prior to transition, and V increases from zero to V_0 in t_0 seconds. V_0' is, of course, that voltage which produces a synchrotron stable region of half width $\sqrt{2}\rho$.

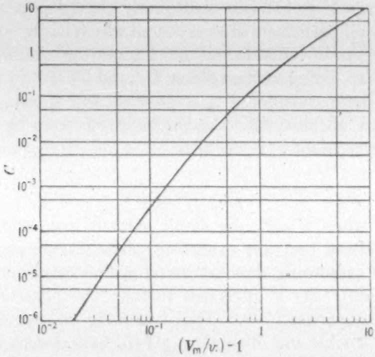


Figure 1. Solution of the equation $[\cos\phi_{s1}-(\frac{1}{2}\pi-\phi_{s1})\sin\phi_{s1}]^2/\tan\phi_{s1}=C$, for a range of values of C, $(\sin\phi_{s1}=v_1/V_m)$.

We may evaluate ρ by extrapolating from betatron injection, assuming that the initial spread of orbits, $2\rho_i$, is one half of the available orbit space, and that these orbits contract inversely as the magnetic field, i.e. as βE (see I). We then obtain

$$V_0' \simeq \pi(1-n)\frac{(\beta_i E_i)^2}{E_T}\left(\frac{\rho_i}{r}\right)^2. \qquad \ldots\ldots(4)$$

Further, if V increases approximately linearly with time, (2) becomes

$$t_m \gg \tfrac{1}{2}\left(\frac{V_m}{V_0}\right)^{3/2}\left[\frac{\pi E_T(1-n)}{w^2 V_m}\right]^{1/2} = \tau, \qquad \ldots\ldots(5)$$

where $V=V_m$ when $t=t_m$.

If (5) is satisfied for $V_0=V_0'$, we obtain the minimum value for V_m from (1). For smaller values of the R.F. rise time, V_0 will be greater than V_0' and V_m will be correspondingly increased.

The above equations may be applied to any given accelerator as follows: Put $V_0=V_0'$ and calculate C, which in this case is given by

$$C = \pi(1-n)\frac{(\beta_i E_i)^2}{v_1 E_1}\left(\frac{\rho_i}{r}\right)^2. \qquad \ldots\ldots(6)$$

Determine V_{m} from (1), or from Figure 1, and τ from (5). It should be noted that C, and thus V_{m}, has no explicit dependence on E_{T}. It may be that τ, as so determined, is too large to enable the inequality $t_{\mathrm{m}} \gg \tau$ to be easily satisfied, in which case the process is repeated with new values of V_0 ($>V_0'$) until (5) is satisfied by an acceptable value of t_{m}.

If the magnetic guide field remains substantially constant during the period t_{m} seconds, the maximum radial oscillation amplitude attained is given approximately by

$$\rho_{\mathrm{m}} = r\left(\frac{V_{\mathrm{m}}}{V_0}\right)^{1/4}\left(\frac{2V_0}{\pi(1-n)E_{\mathrm{T}}}\right)^{1/2}. \qquad \ldots\ldots(7)$$

Thus a further condition for efficient transition is that ρ_{m} should be less than the minimum clearance between the equilibrium orbit at transition and the radial limits of the orbit space. This condition imposes a lower limit on E_{T}. It should be noted that if the magnetic guide field increases substantially during the interval t_{m}, ρ_{m} will be less than the value given by (7) due to the additional phase damping.

§ 3. NUMERICAL EXAMPLES

(i) *Oxford University* 140 MeV. *Synchrotron*

The parameters of this accelerator are: $n=0{\cdot}65$, $E_{\mathrm{i}}=6\times10^5$ ev., $E_{\mathrm{T}}=2{\cdot}5\times10^6$ ev., $\beta_{\mathrm{i}}=0{\cdot}55$, $v_1=375$ volts, $\rho_{\mathrm{i}}\simeq 2{\cdot}0$ cm., $r=48$ cm.

From (4) we obtain $V_0'=80$ volts, and, putting $V_0=V_0'$, we have from (6), $C=5{\cdot}5\times10^5/E_1$. Using this value for C, V_{m} has been estimated for conservative values of E_1/E_{T}. The results are:

E_1/E_{T}	1	2	3	5
V_{m}/v_1	2·0	1·7	1·6	1·4

If $E_1/E_{\mathrm{T}}=3$, the condition for complete capture of the betatron beam becomes $V_{\mathrm{m}}=600$ volts, $t_{\mathrm{m}}\gg 2\,\mu$sec. From (7) we find $\rho_{\mathrm{m}}=0{\cdot}6$ cm., which is substantially less than the clearance between the equilibrium orbit and the radial limits of the orbit space.

(ii) *Glasgow University* 375 MeV. *Synchrotron*

For this machine we have (Wilkins 1950), $n=0{\cdot}7$, $v_1=2{,}500$ volts, $\rho_{\mathrm{i}}=3$ cm., $r=125$ cm. Injection conditions $E_{\mathrm{i}}=6\times10^5$ ev., $\beta_{\mathrm{i}}=0{\cdot}55$, will be assumed, and we will put $E_1=3\times10^7$ ev. This value for E_1 was chosen after careful consideration of the graph of v plotted against magnetic guide field given by Wilkins; it is not critical since even if it is in error by a factor of two, the resultant error in the estimated value of V_{m} will be only a few per cent.

From the foregoing we get $V_0'=5{\cdot}9\times10^7/E_{\mathrm{T}}$, $C=7{\cdot}9\times10^{-4}V_0/V_0'$, $\rho_{\mathrm{m}}=1{\cdot}13\times10^5 E_{\mathrm{T}}^{-3/4}(V_0V_{\mathrm{m}}/V_0'v_1)^{1/4}$, $\tau=4{\cdot}45\times10^{-21}V_{\mathrm{m}}E_{\mathrm{T}}^2(V_0'/V_0)^{3/2}$.

(*a*) $V_0=V_0'$.

The condition for 100% transition efficiency is $V_{\mathrm{m}}=1{\cdot}13v_1=2{,}820$ volts, $t_{\mathrm{m}}\gg\tau$. The following are the values of τ and ρ_{m} for various values of E_{T}:

E_{T} (MeV.)	2	2·5	5	8
τ (μsec.)	50	79	315	805
ρ_{m} (cm.)	2·2	1·85	1·1	0·8

If the electron gun structure extends to a radius of 118·7 cm. (as given by Wilkins), the clearance between equilibrium orbit and gun at $E_{\mathrm{T}}=2$ MeV. is just 2·2 cm. Collision losses will thus be absent provided $E_{\mathrm{T}}>2$ MeV.

(*b*) $V_0 > V_0'$.

Since the above values of τ are rather excessive for a magnet excited at a frequency of 50 c/s., the transition conditions have been determined for a range of values of V_0/V_0' and of E_T. These results are plotted in Figure 2, in which the points enclosed by circles indicate the limits beyond which collisions with the electron gun will occur.

For given values of V_m and E_T, Figure 2 gives the restriction on the time of rise of the R.F. voltage ($t_m \gg \tau$) which must be satisfied for complete capture of the electron beam. For instance, with $V_m = 1{\cdot}3 v_1 = 3{,}250$ volts, $E_T = 2{\cdot}5$ MeV., we obtain $\tau = 2{\cdot}3$ μsec., $\rho_m = 3{\cdot}5$ cm. In this case electrons with maximum amplitude oscillations will just clear the electron gun. With $V_m = 1{\cdot}4 v_1 = 3{,}500$ volts, $E_T = 3$ MeV., we get $\tau = 1$ μsec. and $\rho_m = 3{\cdot}1$ cm. In this case the clearance from the electron gun is 4·3 cm.

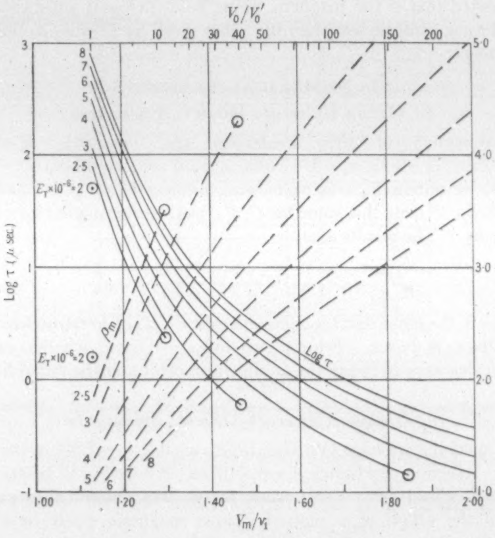


Figure 2. ρ_m and τ as a function of V_m/v_1 and E_T, for the Glasgow University 375 MeV. synchrotron. The points enclosed by circles indicate the limits beyond which collisions with the boundaries of the orbit space may be expected.

§ 4. CONCLUSION

The above treatment allows determination of the R.F. voltage and electron energy required for efficient transition from the betatron to the synchrotron phase, taking account of the finite rate of rise of this voltage. The results are subject to the approximation of I which arises from assuming that the formulae describing the motion of electrons executing small angle phase oscillations are still applicable even when large angle oscillations are involved. For the reasons given in I this is not believed to be a serious source of error.

Perhaps the results derived for the Glasgow synchrotron are the most striking in showing that in this energy range the R.F. voltage needed to capture and retain effectively 100% of the electron beam is not greatly in excess of the peak voltage per revolution required to maintain an electron on the equilibrium orbit.

The avoidance of collision losses in this accelerator merely requires that the electrons should have reached almost the velocity of light prior to transition ($E_T > 2{\cdot}5$ MEV., say). This allows considerable flexibility in betatron flux bar design.

Figure 2 shows that the shorter the R.F. rise time t_m, the larger the peak voltage V_m must be, but that V_m is only a slowly varying function of t_m. The theory does not deduce the variation in efficiency as t_m is reduced towards the value τ. It would decrease, but, at least in some cases, the drop in efficiency may be tolerable (see Appendix).

The above results may be compared with those of other authors. The theory of Goward (1949), which assumes instantaneous application of the full R.F. voltage, requires that $V_m = 2v_1$ for 50% and $V_m = 3{\cdot}3v_1$ for 65% transition efficiency, respectively. Wilkins (1950) applies the theory of Goward to the Glasgow synchrotron. Taking into account the phase oscillation damping which occurs between transition and complete betatron saturation, he finds the optimum condition for 70% efficiency to be $V_m = 5{,}000$ volts and $E_T = 5$ MEV.

APPENDIX

De Packh and Birnbaum (1948) have considered the transition problem assuming a radio-frequency voltage varying as t^s, where s is a constant. By joining a numerical solution of the phase equation, obtained for small values of t, to the adiabatic solution, they have evaluated the maximum radial oscillation amplitudes ρ_m reached in a synchrotron having $r = 300$ cm. and in which V rises from zero to 135,000 volts while the magnetic field increases from 14 to 56 oersted. The peak electron energy is 600 MEV. and the excitation frequency is 60 c/s. Their problem differs from the one considered here, since these authors presumably assume a zero spread of betatron instantaneous orbits prior to $t = 0$. This corresponds most nearly to the case considered above for which $V_0 \gg V_0'$.

From the information given by de Packh and Birnbaum we can calculate V_0 and hence obtain the following approximate expression for the left-hand side of the inequality (2):

$$\frac{300(1+3\alpha)^{3/2+3/s}}{16^{1/s}\alpha^{1+s/2}}\left(\frac{\rho_m}{\rho_{m0}}\right)^{2+4/s} = \gamma \text{ (say)}, \quad s \leqslant 2{\cdot}25,$$

where $\alpha = s/(9-3s)$, and $\rho_m = \rho_{m0}$ when $s = 0$ (i.e. when the R.F. voltage is instantaneously established at a magnetic field of 56 oersted).

The values of γ are given in the Table.

s	0	0·5	1·0	2·0
ρ_m (cm.)	76·8	49·8	22·5	10·9
γ	∞	1·9	1·1	1·9

From these data we would expect the condition $\gamma \gg 1$ not to be particularly strict, especially when $V_0 \gg V_0'$, and to be most easily satisfied when V is approximately a linear function of t, (i.e. $s = 1$), at least for small values of t ($t < t_0$).

REFERENCES

GOWARD, F. K., 1949, *Proc. Phys. Soc.* A, **62**, 617.

KAISER, T. R., 1950, *Proc. Phys. Soc.* A, **63**, 52.

DE PACKH, D. C., and BIRNBAUM, M., 1948, *J. Appl. Phys.*, **19**, 795.

WILKINS, J. J., 1950, *Phil. Mag.*, **41**, 34.

The Measurement of Fluctuation Noise by Diode and Anode-Bend Voltmeters

By R. E. BURGESS

Radio Research Station, Slough, Bucks.

Communicated by R. L. Smith-Rose; MS. received 11*th December* 1950

ABSTRACT. The response of a diode voltmeter to a c.w. sine-wave signal and to noise is analysed for three typical diode characteristics : discontinuous linear, discontinuous parabolic and exponential. In the first type the indication is proportional to the r.m.s. value of the input voltage but the constant of proportionality is different for noise and for c.w. and depends upon the ratio of the load resistance to diode resistance. For small input signals the curvature of the diode characteristic is important and the voltmeter tends to a square-law behaviour with the result that c.w., noise or a mixture of c.w. and noise can all be measured in terms of a single calibration.

The input conductance of a diode voltmeter is in general greater for noise than for c.w. It can be very much greater in practical conditions and this represents a possible source of error in the measurement of noise.

The response of the anode-bend voltmeter to c.w. and to noise is considered for a discontinuous parabolic and for an exponential characteristic. The departure of the c.w. calibration from a square law and the difference between c.w. and noise calibrations are evaluated in terms of the characteristics.

§1. INTRODUCTION

Measurements of valve and circuit noise and of other types of fluctuation noise (e.g. cosmic noise) are made with a variety of types of instrument. In those investigations in which high absolute accuracy is required, as for example in the determination of the electronic charge or of Boltzmann's constant, it has been customary to use a thermocouple, since its indication is strictly square-law and thus r.m.s. values of the noise current (or voltage) can be determined accurately.

In investigations where high absolute accuracy is not called for, as for instance in relative measurements of valve and circuit noise, it has been found more convenient to use a valve voltmeter on account of the delicacy and inertia of the thermocouple. The valve voltmeter is sometimes used on the assumption that its c.w. calibration also applies to noise, but this is only true if the characteristic is accurately parabolic over the working range, a condition which may be difficult to satisfy if a wide range of voltages is to be measured.

The difficulties which arise in the measurement of fluctuation noise are due rather to the characteristics of the noise than to departures of the rectifier from the ideal. It is known that although a fluctuation noise voltage has a definite r.m.s. value it has no definite peak value; the instantaneous voltage and amplitude are distributed according to certain statistical laws which depend upon the essential randomness of the elementary events responsible for the noise. Thus relatively large values of amplitude correspond to relatively infrequent conditions in which the noise components are largely cophasal and are therefore of small probability.

It is assumed in the present note that the noise applied to the voltmeter emerges from an amplifier whose bandwidth is small compared with the mean frequency. The emergent noise therefore has an oscillatory form whose mean frequency is equal to that of the amplifier and whose amplitude varies from moment to moment at a rate which is directly dependent on the bandwidth.

The probability distribution of the instantaneous voltage v of fluctuation noise of r.m.s. voltage E obeys the Gaussian law

$$P(v) = [1/\sqrt{(2\pi)}E] \exp(-v^2/2E^2). \qquad \ldots\ldots(1.1)$$

The instantaneous amplitude A of the noise envelope obeys the Rayleigh law

$$P(A) = (A/E^2) \exp(-A^2/2E^2). \qquad \ldots\ldots(1.2)$$

A mixture of a sine-wave voltage of amplitude S and fluctuation noise of r.m.s. voltage E will have for the probability distribution of its instantaneous voltage v

$$P(v) = \frac{1}{\pi\sqrt{(2\pi)}E} \int_0^\pi \exp\left(-\frac{(v - S\cos\phi)^2}{2E^2}\right) d\phi \qquad \ldots\ldots(1.3)$$

which does not appear to be expressible in a closed form involving tabulated functions.

These probability distributions permit the calculation of the response of a non-linear system to fluctuation noise in the absence or presence of a c.w. signal.

The types of valve voltmeter discussed are the diode voltmeter and the anode-bend voltmeter. An essential difference between the two types is that in the diode voltmeter the rectified voltage appearing across the load resistor displaces the operating point of the valve whereas in an anode-bend voltmeter with negligible anode and cathode circuit resistances this does not occur. Certain types of valve voltmeter fall intermediately between these two types, e.g. the reflex instrument in which the increase of mean anode current produced by the applied R.F. voltage increases the cathode bias so extending the range and tending to linearize the calibration. The leaky-grid voltmeter can be regarded as a diode voltmeter of the exponential type discussed below followed by a linear D.C. amplifier.

§ 2. DIODE VOLTMETER

2.1. *Introduction*

In the analysis of diode rectification now presented *it is assumed that the time constant RC of the load circuit is large compared with the reciprocal of the amplifier bandwidth* (Figure 1) so that the rectifier voltage V is sensibly constant.

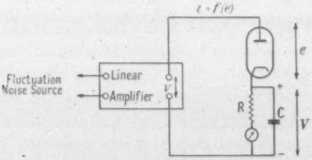


Figure 1. The rectifier circuit.

In a radio receiver (as distinct from a valve voltmeter) this condition does not apply, for the time constant of the diode load circuit is chosen to be small compared with the reciprocal of the bandwidth of the preceding circuits in order that the modulation components of the signals may be faithfully extracted. This type of detector may be termed an 'envelope follower' and its properties are discussed in a paper in course of publication by the author.

The general conditions for rectification of a c.w. sinusoidal signal and of fluctuation noise can now be deduced. In the case of c.w. rectification of a signal $v = S\cos\omega t$, the instantaneous potential difference across the rectifier is $e = v - V = S\cos\omega t - V$. The current–voltage characteristic of the rectifier is assumed to be represented by $i = f(e)$. The steady state condition corresponds to equality of the rate of discharge of C through R to the rate of charge through the diode and can be expressed as follows:

$$\frac{V}{R} = \frac{1}{2\pi}\int_0^{2\pi} f(v - V)\,d(\omega t). \qquad \ldots\ldots(2.1)$$

In the case of the rectification of fluctuation noise this equation must be modified to the completely general equation

$$\frac{V}{R} = \int_{-\infty}^{+\infty} f(v - V)P(v)\,dv. \qquad \ldots\ldots(2.2)$$

If the detector is conducting only for anode voltages above a certain value, the lower limit of the integral can be correspondingly modified.

In the following sections the linear, exponential and parabolic characteristics are discussed and are taken as being applicable with fair accuracy to the case of large, small and intermediate signal rectification respectively.

The rectification is conveniently measured by the coefficients Y and Z for c.w. and noise respectively which are defined as the ratio of the rectified voltage due to the application of the c.w. or noise to $\sqrt{2}$ times the r.m.s. value σ of the applied voltage. In the case of c.w. it is seen that Y is the ratio of the rectified voltage to the amplitude S of the applied voltage and therefore corresponds to the usual definition of the efficiency of rectification.

2.2. *Discontinuous Linear Characteristic*

The characteristic considered corresponds to zero conductance for negative anode voltages and a constant conductance g for positive anode voltages: $i = 0,\ e < 0;\ i = ge,\ e > 0$. This characteristic is applicable to a fair approximation for relatively large applied voltages.

The c.w. rectification coefficient $V/S = Y$ is given implicitly by

$$gR = \pi Y/[(1 - Y^2)^{1/2} - Y\cos^{-1} Y] \qquad \ldots\ldots(2.3)$$

and the lower curve in Figure 2 shows Y as a function of gR.

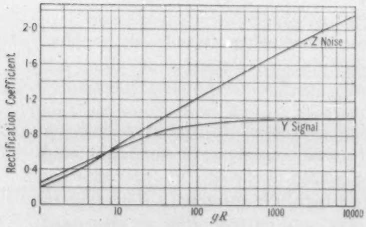


Figure 2. Rectification coefficient for a c.w. signal (Y) and noise (Z) with a linear characteristic.

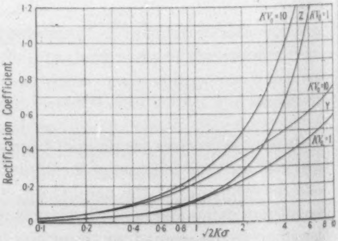


Figure 3. Rectification coefficient for a c.w. signa (Y) and noise (Z) with an exponential characteristic

If fluctuation noise of r.m.s. voltage E be applied to the rectifier the rectified voltage V is given by $V/R = g\int_V^\infty (v-V)P(v)\,dv$ and thus the rectification coefficient Z for noise is given implicitly by

$$\frac{1}{gR} = \frac{1}{Z\sqrt{\pi}}\int_Z^\infty (u-Z)\exp(-u^2)\,du = \frac{\exp(-Z^2)}{2Z\sqrt{\pi}} - \tfrac{1}{2}(\operatorname{erfc} Z) \quad \ldots\ldots(2.4)$$

where erfc Z is the complementary error function defined by

$$\operatorname{erfc} Z = (2/\sqrt{\pi})\int_Z^\infty \exp(-u^2)\,du = 1 - \operatorname{erf} Z.$$

The upper curve in Figure 2 shows Z as a function of gR.

It is seen that for small values of gR the noise coefficient Z is less than the C.W. coefficient Y. From equations (2.3) and (2.4) it is found that, for gR small, $Y \simeq gR/\pi$; $Z \simeq gR/2\sqrt{\pi} = \frac{1}{2}\sqrt{\pi}Y$. The limiting ratio of $\frac{1}{2}\sqrt{\pi}$ is the value to be expected, since it is the ratio of the average values of noise and C.W. of the same r.m.s. value.

For large values of gR it is seen that whereas Y tends slowly to unity (peak rectification) Z increases steadily. It is found from equation (2.4) using the asymptotic expansion of the error function that, for gR large, $gR \simeq 4\sqrt{\pi}Z^3 \exp Z^2$ which shows that Z is a slowly increasing function of gR.

It will be noted that in the case of noise the rectified voltage $V = Z\sqrt{2}E$ can appreciably exceed the average amplitude and thus the diode voltmeter does not measure the average of the envelope, but a pseudo-peak value which in practice may be three to four times the r.m.s. voltage.

An important feature of the linear rectifier is that the rectification coefficients Y and Z are independent of the applied voltage and in fact only functions of gR. Thus such an ideal rectifier gives an indication which is proportional to the r.m.s. value of the applied voltage (whether C.W. or noise). In practice departure from proportionality is found to occur at the smaller voltages (say less than 5 V. r.m.s.) due to the rectifier no longer being linear, i.e. due to the decrease of g with a decrease of applied voltage.

2.3. *Alternative Approach in Terms of Noise Amplitude*

In a previous paper (Burgess 1948) the author has shown generally that if an ideal linear diode is supplied with an input R.F. noise voltage the rectified voltage V is such that

$$\overline{\tan\theta - \theta} = \pi/gR \quad \ldots\ldots(2.5)$$

where g is the diode conductance, R the load resistance and 2θ the instantaneous conduction angle during a cycle, the average (denoted by the bar) to be taken over all the possible cases arising from the amplitude variations of the noise voltage. It is assumed that the discharge time constant is large compared with the charge time constant ($gR \gg 1$) and is also sufficiently large for the rectified voltage to be substantially constant with time.

For a C.W. voltage of amplitude S the angle θ is constant and given by $\theta = \cos^{-1} V/S$ which leads to the usual formula (equation (2.3)) for C.W. rectification. For recurrent rectangular R.F. impulses of duration T and spacing T_0 the conduction angle is say $2\theta_1$ during an impulse and zero during the spaces whence $(\tan\theta_1 - \theta_1)T/(T+T_0) = \pi/gR$.

In the case of fluctuation noise the instantaneous amplitude A has a Rayleigh distribution $P(A) = (A/E^2)\exp(-A^2/2E^2)$.

The conduction semi-angle θ for a cycle of the noise wave of amplitude A is given by $V=A\cos\theta$ and thus the probability distribution of θ is

$$P(\theta)=\frac{V^2\sec^2\theta\tan\theta}{E^2}\exp\left(-\frac{V^2\sec^2\theta}{2E^2}\right). \qquad \ldots\ldots(2.6)$$

The average value of $\tan\theta-\theta$ is thus

$$\overline{\tan\theta-\theta}=\frac{V^2}{E^2}\int_0^{\pi/2}(\tan\theta-\theta)\sec^2\theta\tan\theta\exp\left(-\frac{V^2\sec^2\theta}{2E^2}\right)d\theta$$
$$=2Z^2\int_0^{\infty}u(u-\tan^{-1}u)\exp\{-Z^2(1+u^2)\}\,du,$$

where $u=\tan\theta$ and $Z=V/\sqrt{2}E$ as before. Hence from (2.5)

$$\frac{\pi}{2gRZ^2}=\frac{\sqrt{\pi}\exp(-Z^2)}{4Z^3}-\int_0^{\infty}u\tan^{-1}u\exp[-Z^2(1+u^2)]\,du.$$

Now $\int_0^{\infty}u\tan^{-1}u\exp[-Z^2(1+u^2)]\,du=\frac{1}{2Z^2}\int_0^{\infty}\frac{\exp[-Z^2(1+u^2)]}{1+u^2}du=\frac{\pi}{4Z^2}\operatorname{erfc}Z.$

Thus $1/gR=\{\exp(-Z^2)\}/2\sqrt{\pi}Z-\frac{1}{2}\operatorname{erfc}Z$ which is equation (2.4) above.

2.4. *Exponential Characteristic*

When very small signals (say less than 0·5 v. r.m.s.) are applied to the diode rectifier, the relevant portion of the characteristic is the initial current region which is exponential and expressed by

$$i=i_0\exp(Ke) \qquad \ldots\ldots(2.7)$$

in which i_0 and K are constants; for modern valves with normal cathode temperatures, $K\simeq 10\,\text{volt}^{-1}$. The exponential characteristic is also applicable to the leaky-grid voltmeter.

A feature of the exponential rectifier is the initial standing current which exists in the absence of an applied voltage. Thus the rectification coefficients are given in terms of $(V-V_0)$ where V_0 is the standing voltage across the load resistance in the absence of a signal or noise and V is the voltage on application of these.

From equation (2.7) it follows that V_0 is given by the standing-current equation

$$KV_0\exp(KV_0)=KRi_0. \qquad \ldots\ldots(2.8)$$

In practice, KV_0 usually lies between 1 and 10 (i.e. V_0 lies between about 0·1 and 1 volt).

If a C.W. voltage of amplitude S is applied to the rectifier, the resulting potential difference V across R is given by

$$V=\frac{R}{\pi}\int_0^{\pi}i_0\exp(KS\cos\phi-KV)\,d\phi=i_0R\exp(-KV)I_0(KS) \qquad \ldots\ldots(2.9)$$

in which I_0 is the modified Bessel function of the first kind and order zero. From the last two equations it is found that the rectification coefficient Y for a C.W. signal of amplitude S is given by

$$(YKS+KV_0)\exp(YKS)=KV_0I_0(KS). \qquad \ldots\ldots(2.10)$$

This relation has already been derived by Aiken (1938).

In Figure 3, Y is shown as a function of KS for $KV_0=1$ and 10 in the two lower curves. It is seen that the greater the value of KV_0 (i.e. the greater the standing voltage) the greater is the efficiency of rectification.

If fluctuation noise of r.m.s. voltage E is applied to the rectifier, the voltage V appearing across the load resistance is given by

$$V=R\int_{-\infty}^{+\infty} i_0 \exp K(v-V)P(v)\,dv=\frac{Ri_0}{\sqrt{\pi}}\exp(-KV)\int_{-\infty}^{+\infty}\exp(\sqrt{2}\,KEu-u^2)\,du$$
$$=Ri_0\exp(-KV)\exp(\tfrac{1}{2}K^2E^2) \qquad \ldots\ldots(2.11)$$

and thus the rectification coefficient Z for noise is given by

$$(Z\sqrt{2}\,KE+KV_0)\exp(\sqrt{2}\,ZKE)=KV_0\exp(\tfrac{1}{2}K^2E^2). \qquad \ldots\ldots(2.12)$$

The upper curves in Figure 3 give Z as a function of $\sqrt{2}\,KE$ for $KV_0=1$ and 10.

If a mixture of a sine wave of amplitude S and fluctuation noise of r.m.s. voltage E be applied to the exponential diode rectifier it is seen from equation (1.3) that the voltage V appearing across the load resistance is given by

$$V=\frac{Ri_0}{\pi\sqrt{(2\pi)}E}\int_0^{\pi}d\phi\int_{-\infty}^{+\infty}\exp\left[K(v-V)-\frac{(v-S\cos\phi)^2}{2E^2}\right]dv$$
$$=Ri_0\exp(-KV+\tfrac{1}{2}K^2E^2)I_0(KS). \qquad \ldots\ldots(2.13)$$

Equations (2.10) and (2.12) are special cases of this.

If the input voltage is small (KE and $KS \ll 1$),

$$V\exp KV \simeq Ri_0[1+\tfrac{1}{2}K^2(E^2+\tfrac{1}{2}S^2)]$$

and writing $V\exp KV \simeq V_0\exp KV_0+[(1+KV_0)\exp KV_0](V-V_0)$ and $\overline{v^2}\equiv E^2+\frac{1}{2}S^2$ we obtain in conjunction with (2.8) $V-V_0=\frac{1}{2}K^2V_0\overline{v^2}/(1+KV_0)$.

This shows that the indication is proportional to the mean square input voltage whether this be C.W., noise or a mixture of C.W. and noise. Hence a calibration made on C.W. will be generally applicable.

The behaviour of the coefficients Y and Z for high values of applied voltage is not of practical interest, since the rectifier will no longer be exponential in these conditions. In order to cover the intermediate range of applied voltage in which the diode characteristic is neither rising as rapidly as the exponential function nor is sufficiently linear to be considered to have a constant conductance, the case of the parabolic characteristic is discussed in the next section.

2.5. *Discontinuous Parabolic Characteristic*

For intermediate values of applied voltage, the diode characteristic can frequently be represented by a discontinuous parabolic characteristic of the form $i=\alpha e^2$, for $e>0$; $i=0$, for $e<0$.

If a C.W. signal of amplitude S be applied to the rectifier, the rectification coefficient Y is given implicitly by

$$S\alpha R=2\pi Y/[(1+2Y^2)\cos^{-1}Y-3Y(1-Y^2)^{1/2}]. \qquad \ldots\ldots(2.14)$$

The lower curve in Figure 4 shows Y as a function of $S\alpha R$ and the usual slow rise to unity is exhibited.

The appropriate integral equation for the rectification of noise of r.m.s. voltage E is $V/R=\alpha\int_V^{\infty}(v-V)^2P(v)\,dv$. The noise detection coefficient is given implicitly by

$$\frac{1}{\sqrt{2}\,E\alpha R}=\frac{1}{Z\sqrt{\pi}}\int_Z^{\infty}(u-Z)^2\exp(-u^2)\,.\,du=\frac{1+2Z^2}{4Z}(1-\operatorname{erf}Z)-\frac{\exp(-Z^2)}{2\sqrt{\pi}}.$$

The upper curve in Figure 4 shows Z as a function of $\sqrt{2}^{7}E\alpha R$; the Y and Z curves for the parabolic characteristic are seen to resemble those for the linear diode with Z rising uniformly but slowly as the load resistance R is increased.

Since Y and Z are sensibly constant over a wide range of input voltage it can be inferred that the parabolic rectifier gives a linear indication of the r.m.s. value of the applied signal or noise over a fair range of voltage. Only for very small applied voltages below the values covered by Figure 4 does the rectifier become square-law, for it is seen from equation (2.24) that, for $Y \ll 1, Y \simeq \frac{1}{4}S\alpha R$.

In this region it is found that the noise rectification coefficient is also given by $Z \simeq E\alpha R/2\sqrt{2}$; thus for both signal and noise the rectification is square-law for very small applied voltages and the same calibration applies to both, namely $V = \frac{1}{2}\alpha R \times$ (mean square input voltage). However, in practice the value of the voltage necessary for the rectifier to behave as a square-law instrument is probably so small that the characteristic is exponential rather than parabolic over the relevant portion.

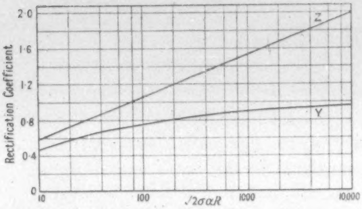


Figure 4. Rectification coefficient for a C.W. signal (Y) and noise (Z) with a parabolic characteristic.

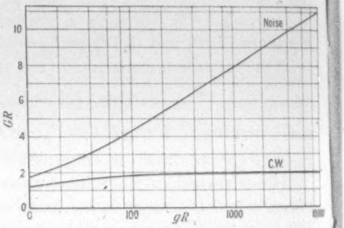


Figure 5. Effective input conductance of a linear rectifier for C.W. and for noise.

2.6. *Power Consumed by Diode Rectifier*

A diode rectifier consumes power in both the diode and the load resistance. In the case of a C.W. signal of amplitude S the average power consumed is given by

$$W = \frac{1}{\pi}\int_0^{\pi} S\cos\phi f(S\cos\phi - v)d\phi,$$

while in the case of noise the average power is given by

$$W = \int_{-\infty}^{\infty} Vf(v-V)P(v)\,dv.$$

These expressions may be used to calculate the equivalent input conductance G of the rectifier circuit which is given by

$$\text{C.W. signal:}\quad G = \frac{2W}{S^2} = \frac{2}{\pi S}\int_0^{\pi}\cos\phi f(S\cos\phi - V)\,d\phi \qquad \ldots\ldots(2.15)$$

$$\text{Noise:}\quad G = \frac{W}{E^2} = \frac{1}{\sqrt{(2\pi)}E^3}\int_{-\infty}^{+\infty} vf(v-V)\exp\left(-\frac{v^2}{2E^2}\right)dv \qquad \ldots\ldots(2.16)$$

$$= \frac{1}{\sqrt{(2\pi)}E}\int_{-\infty}^{+\infty} f'(v-V)\exp\left(-\frac{v^2}{2E^2}\right)dv. \qquad \ldots\ldots(2.16a)$$

That these conductances are in general different may be seen by considering the duration of the conduction periods in the two cases. When a C.W. signal is applied to a rectifier with a long time constant, the rectified voltage V builds up to the equilibrium value corresponding to the amplitude of the signal, and in the case of the linear rectifier the conduction angle depends only upon the product gR and is independent of the amplitude of the signal. If noise is applied to a linear rectifier, however, the conditions are essentially different, since the amplitude varies from moment to moment, but the rectified voltage V cannot readjust itself at each cycle. The conduction angle may therefore approach 180° on the larger peaks of amplitude, but will be zero for all amplitudes less than V. Equation (2.16 *a*) shows in fact that the effective input conductance for noise is equal to the average value of the diode conductance f'.

The effective input conductance which a linear rectifier presents to a tuned circuit exciting it with fluctuation noise is given by

$$G = \frac{g}{E^2}\int_V^\infty v(v-V)P(v)\,dv = \frac{g}{2}(\operatorname{erfc} Z)$$

in which Z has the value given in Figure 2. Thus for gR small G tends to $\frac{1}{2}g$ (as for C.W.) while for gR large G tends to $2Z^2/R$, which corresponds to the power expended in the load resistance. Figure 5 shows GR as a function of gR for rectification of C.W. and of noise.

It is seen that the damping effect of a rectifier has a value for noise which is not the same as for C.W. and this effect is a possible source of error in measurement of noise when calibration is carried out on C.W. or on a type of noise different from that being measured.

For the exponential diode the conductance for a noise input is

$$G = \frac{1}{\sqrt{(2\pi)}E^3}\int_{-\infty}^{\infty} v i_0 \exp\left[K(v-V) - \frac{v^2}{2E^2}\right]dv = Ki_0\exp(-KV + \tfrac{1}{2}K^2E^2) = \frac{KV}{R}$$

by virtue of equation (2.11). Thus the effective input conductance is simply K times the current flowing in the load resistor. For very small signal inputs where V is sensibly equal to V_0 the input conductance is KV_0/R, that is proportional to the initial diode current.

For C.W. the input conductance is given by

$$G = \frac{2i_0}{\pi S}\int_0^\pi \cos\phi \exp[K(S\cos\phi - V)]\,d\phi = \frac{2i_0}{S}\exp(-KV)I_1(KS)$$

$$= \frac{KV}{R}\,\frac{2I_1(KS)}{KSI_0(KS)}$$

by virtue of equation (2.10). This result agrees with that of Aiken (1938). Now the second term tends to unity when KS is small and the input conductance is therefore the same, KV/R, for noise and for C.W. at low input levels. For KS large I_0 and I_1 tend to equality as do V and S, and the input conductance for C.W. becomes $2/R$. The conductance for noise is always greater than for C.W. with a given rectified voltage V.

§ 3. THE ANODE-BEND VOLTMETER

3.1. *General Analysis*

The anode-bend voltmeter which is more robust, has less inertia than a thermocouple and a lower input conductance than a diode voltmeter, is commonly used for measuring noise and mixtures of signal and noise, for, by suitable choice of valve and operating conditions, the calibration can be made to obey a square law fairly accurately. The object of the present section is to analyse the calibrations of an anode-bend voltmeter for C.W., for noise and for a mixture of C.W. and noise assuming either a discontinuous parabolic or an exponential characteristic.

In practice the grid bias would be adjusted to give a C.W. calibration which is closely square-law over an appreciable range of input voltages; it is also necessary to avoid grid current in order that the input conductance shall be small.

The anode circuit, which will contain the indicating meter with suitable provision for backing-off the initial current, is assumed to be of low impedance for both radio-frequency and direct current, so that the change of mean anode potential and its fluctuations on applying a noise voltage can be neglected in calculating the change of anode current.

Let the anode current–grid voltage characteristic be $i = f(v + v_0)$ with $i_0 = f(v_0)$ for the standing anode current.

When C.W. of amplitude S is applied, the mean anode current is

$$\bar{i} = \frac{1}{\pi} \int_0^\pi f(S \cos\phi + v_0)\, d\phi. \qquad \ldots\ldots(3.1)$$

When noise of r.m.s. value E is applied the mean anode current is given by

$$\bar{i} = \frac{1}{\sqrt{(2\pi)}E} \int_{-\infty}^{\infty} f(v + v_0) \exp(-v^2/2E^2)\, dv. \qquad \ldots\ldots(3.2)$$

It is seen that the average anode current obeys the differential equation $\partial\bar{i}/\partial(E^2) = \frac{1}{2}\partial^2\bar{i}/\partial v_0^2$ thus relating the response to noise and the curvature of the characteristic in the presence of the noise. For small noise inputs, the simple result $\bar{i} \simeq i_0 + E^2 \partial\bar{i}/\partial(E^2) = i_0 + \frac{1}{2}E^2 f''(v_0)$ is obtained corresponding to the expansion of $\bar{i}$ in terms up to E^2.

The differential equation shows that the curvature term $f''(v_0)$ is the significant quantity and determines the rectification process at low levels. For with *any* form of voltage $\bar{i} = \overline{f(v + v_0)} = f(v_0) + \bar{v} f'(v_0) + \frac{1}{2}\overline{v^2} f''(v_0) + \ldots$ and square-law rectification holds at low levels for C.W., for noise and for a mixture of C.W. and noise. The range of input voltage over which the square-law behaviour can be expected depends on the ratio of $f''(v_0)$ to the higher derivatives and the selection of operating conditions is a question of making this ratio as large as possible.

If a mixture of C.W. and noise is applied, the mean current is

$$\bar{i} = \frac{1}{2\pi} \int_0^{2\pi} d\phi \, \frac{1}{\sqrt{(2\pi)}E} \int_{-\infty}^{+\infty} f(S \cos\phi + v + v_0) \exp(-v^2/2E^2)\, dv. \qquad \ldots\ldots(3.3)$$

Although the integrals of equations (3.1) and (3.2) are often expressible in terms of tabulated functions, the integral of equation (3.3) is generally not readily evaluated; the case of the exponential characteristic is an exception since the C.W. and noise factors are then separable.

The anode-bend voltmeter differs from the diode voltmeter in that the rectified voltage does not displace the rectification characteristic. It is this displacement which leads to high peak voltage readings of noise on a diode voltmeter and, analytically, causes the rectified voltage to appear in the integrand and in the limits of integration in the case of a discontinuous characteristic.

3.2. *Discontinuous Parabolic Characteristic*

This characteristic which is of most practical interest is given by

$$i = a(v+v_0)^2 \text{ for } (v+v_0)>0\,; \; i=0 \text{ for } (v+v_0)<0$$

where $-v_0$ is the cut-off voltage relative to the operating point.

It is convenient to introduce the parameter

$$\begin{aligned} x = \frac{\text{r.m.s. input voltage}}{v_0/\sqrt{2}} &= S/v_0 \qquad \text{for c.w.} \\ &= \sqrt{2}E/v_0 \qquad \text{for noise.} \end{aligned}$$

Then for a c.w. input voltage the indication, defined as $y=(\bar{i}-i_0)/i_0$, is given by

$$\begin{aligned} y &= \tfrac{1}{2}x^2 && \text{for } x<1 \\ &= \tfrac{1}{2}x^2+[3(x^2-1)^{1/2}-(x^2+2)\cos^{-1}x^{-1}]/2\pi && \text{for } x>1 \\ &\sim \tfrac{1}{4}x^2+2x/\pi-\tfrac{1}{2} && \text{for } x\gg 1. \end{aligned}$$

The indication is exactly square-law for amplitude less than the cut-off voltage and approximately square-law for larger amplitudes, the calibration constant (y/x^2) falling steadily from $\tfrac{1}{2}$.

When a noise voltage is applied the indication is given by

$$y = \frac{2+x^2}{4}\left(1+\operatorname{erf}\frac{1}{x}\right)+\frac{x}{2\sqrt{\pi}}\exp\left(-\frac{1}{x^2}\right)-1$$

$$\simeq \tfrac{1}{2}x^2 \text{ for } x\ll 1 \text{ and } \sim \tfrac{1}{4}x^2+x/\sqrt{\pi}-\tfrac{1}{2} \text{ for } x\gg 1.$$

The indication on noise is therefore slightly less than that on c.w. of the same r.m.s. voltage. Figure 6(*a*) shows y as a function of x^2 for c.w. and for noise, and the difference is seen to be small; consequently the calibration made with c.w. should be reasonably accurate when applied to noise, the maximum error in the indicated r.m.s. value of a noise voltage when using the c.w. calibration being about 3%. The error in the measurement of a mixture of c.w. and noise should not exceed this and for this reason it has not been thought worth while to carry out the rather laborious complete analysis for the case of a mixture.

It is readily shown that for a discontinuous m-power characteristic, and *for a sufficiently small applied voltage* the calibration is square-law and the same for any type of input voltage, namely $y=\tfrac{1}{2}m(m-1)\overline{v^2}/v_0{}^2$ showing that a calibration of r.m.s. voltage of c.w. will also apply to noise and to a mixture of signal and noise.

3.3. *Exponential Characteristic*

When the input voltage is small and the operating point is near to 'cut-off' the characteristic can usually be represented by an exponential function $i=i_0\exp(Kv)$. The mean anode current for a c.w. input is

$$\bar{i} = \frac{i_0}{2\pi}\int_0^{2\pi}\exp(KS\cos\theta)\,d\theta = i_0 I_0(KS)$$

giving $y=I_0(KS)-1\simeq\tfrac{1}{4}K^2S^2$ for $KS\ll 1$ where I_0 is the Bessel function of the first kind for imaginary argument.

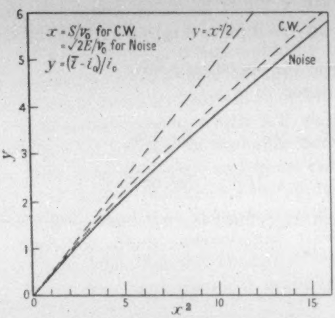


Figure 6 (*a*). Calibration of anode-bend voltmeter for discontinuous parabolic characteristic.

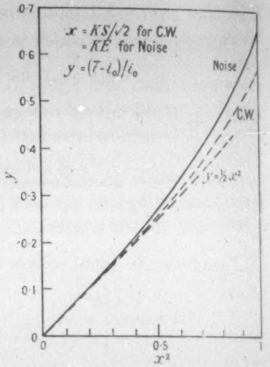


Figure 6 (*b*). Calibration for exponential characteristic.

The mean anode current for a noise input voltage is

$$\bar{i} = \frac{i_0}{\sqrt{(2\pi)}E}\int_{-\infty}^{\infty} \exp\left(Kv - \frac{v^2}{2E^2}\right) dv = i_0 \exp\left(\tfrac{1}{2}K^2E^2\right)$$

giving $y = \exp(\frac{1}{2}K^2E^2) - 1 \simeq \frac{1}{2}K^2E^2$ for $KE \ll 1$.

For a *mixture of noise and* C.W. the integral of equation (3.3) gives

$$\bar{i} = i_0 \exp(\tfrac{1}{2}K^2E^2) I_0(KS)$$

or $$y = \exp(\tfrac{1}{2}K^2E^2) I_0(KS) - 1 \simeq \tfrac{1}{2}K^2(E^2 + \tfrac{1}{2}S^2) \text{ for } K(E^2 + \tfrac{1}{2}S^2)^{1/2} \ll 1.$$

Hence for very small input voltages ($x < 0{\cdot}5$ say) whether C.W., noise or a mixture, the simple square-law behaviour obtains: $y = \frac{1}{2}x^2$, where x is now defined as $K \times$ (r.m.s. value of the applied voltage). This is because the rectification is mainly provided by the quadratic term in the exponential function.

Figure 6 (*b*) shows y as a function of x^2 for the exponential characteristic. The indication on noise is seen to be greater than for C.W. of the same r.m.s. value. However, the dependence for larger input voltages is not of practical interest since the characteristic would begin to depart from the exponential law.

ACKNOWLEDGMENTS

The work described in this paper was carried out as part of the programme of the Radio Research Board to whom it was communicated in confidential reports in 1944 and 1945. It is published by permission of the Director of the National Physical Laboratory and the Director of Radio Research of the Department of Scientific and Industrial Research.

REFERENCES

AIKEN, C. B., 1938, *Proc. Inst. Radio Engrs., N.Y.*, **26**, 859.
BURGESS, R. E., 1948, *J. Instn. Elect. Engrs.*, **95**, Pt. III, 106.

Mechanism of Secondary Ionization in Low-Pressure Breakdown in Hydrogen

By F. LLEWELLYN JONES and D. E. DAVIES

Department of Physics, University College of Swansea

MS. received 31st January 1951

ABSTRACT. The breakdown mechanism was investigated by examining the dependence of the shape of the cathode emission (ω/α, X/p) curves on the deposition of electro-positive or electro-negative atoms on cathodes of nickel, aluminium, silver, copper and molybdenum. X is the electric intensity, and p the gas pressure at about 15° c., and ω is the generalized secondary ionization coefficient which includes processes of cathode emission due to impact of positive ions, of excited atoms, and of photons generated in the ionization current; α is the primary ionization coefficient due to the electrons. Deposition of positive atoms on the cathode produced sharp photoelectric peaks in the curves in the region $X/p \simeq 150$ v./cm. mm. Hg, but ω/α was not greatly affected at higher values of X/p. These results are analysed in terms of the varying effective work function of the surface due to the deposition of atoms. It is concluded that the high photoelectric emission from cathodes of low effective work function was due to low-energy photons generated in the pre-breakdown currents, but that for clean metals the secondary emission was due to impact of positive ions; photo-emission due to any high-energy photons generated in the ionization currents was negligible in comparison.

§ 1. INTRODUCTION

SECONDARY electron emission from the cathode may be produced by any of the three processes: impact of positive ions (γ-effect), of excited atoms (δ'-effect), or of photons (δ-effect) generated in the electron avalanche, since in the present experiments ionization by collision by positive ions (β-effect) is negligible. Thus when α is the primary ionization coefficient due to electrons, the generalized secondary ionization coefficient is given by $\omega = \alpha\gamma + \delta + \delta'$.

Previous experimental investigations (Llewellyn Jones and Davies 1951) of the effects of thin cathode layers of oxides or of metals have demonstrated the very great influence which such films have on the sparking potential of hydrogen at low pressures. It is highly probable that the main result of the deposition of a thin film on the cathode is to change the effective work function of the surface and thereby change the coefficient ω/α of electron emission due to the incidence of photons or of positive ions. The technique which the previous experiments showed to be most effective in removing last traces of oxide layers was that of bombardment by positive ions in a hydrogen glow discharge. It is probable, however, that this procedure also had the effect of producing a layer of hydrogen ions or of atoms on top of whatever metal film was already on the cathode.

In the present paper are described further investigations of the influence of such cathode films in hydrogen with a view to explaining the phenomena, and the results are discussed in relation to the theory of spark discharge mechanism at low pressures. The experimental technique employed was generally similar to that used in the previous experiments. The electrode surfaces were subjected to various treatments such as heating *in vacuo*, bombardment by hydrogen positive ions, and the deposition of metallic films; and the state of the electrodes was observed at intervals during the treatment by measurement of the sparking potential in pure hydrogen. From such data the coefficient ω/α could be found, and its value used as an indication of the state of the surface: the variation of this factor ω/α with X/p can provide information concerning the significant fundamental processes involved in the breakdown.

§2. EXPERIMENTAL PROCEDURE

The electrodes consisted of discs 2·5 cm. in diameter, with bevelled edges, and they were mounted parallel 2 mm. apart by means of a hard glass frame. At the back of each electrode was fitted a small filament of tungsten wire, with connections brought out through separate seals. The electrodes were mounted in pairs in cylinders of hard glass of 4·8 cm. bore, the inner surface of which was screened from the electrode system by a cylinder of pure silver foil fitting tightly inside the envelope. This silver cylinder, which prevented the motion of metal ions or atoms to or from the glass, was connected via a resistance of 500 MΩ to a potential midway between those of the electrodes. Three discharge tubes were made containing the following pairs of electrodes: tube No. 1, silver and molybdenum, tube No. 2, nickel and aluminium, and tube No. 3, copper and nickel. These metals were obtained in the highest state of purity. Hydrogen was generated by the electrolysis of a hot solution of barium hydroxide, and after drying was diffused through the walls of a heated palladium tube, and finally was cooled by liquid air for half an hour before entering the discharge tubes.

The electrodes were outgassed at red heat with the filaments glowing brightly for some hours while the tubes were continuously evacuated. The tubes were then isolated by closing the appropriate stopcocks, and were allowed to stand for about four weeks, after which the pressure was found to be still no more than 10^{-4} mm. Hg. The tubes were given an initial treatment consisting of the admission of pure hydrogen at a pressure of about 10 mm. Hg, baking the envelopes at 400° c. for 12 hours with the electrode filaments glowing, and afterwards evacuating while still hot: this treatment removed most of any oxide layers and left probably only extremely thin oxide films. The state of the electrode surfaces during subsequent treatment was then investigated by taking measurements of sparking potential V_s.

A set of Paschen curves was taken in pure hydrogen with each electrode in turn as cathode. Great care was taken in all these measurements so that only a single breakdown was necessary to measure V_s; measurements were also made in the three tubes in ascending order of V_s. This procedure was adopted in order to minimize any effect on the electrode due to the actual measurements themselves, a procedure previously found to be effective (Llewellyn Jones and Henderson 1939). After this, the tubes were again baked continuously at 400° c. *in vacuo* and there was additional heating of the electrodes by the glowing filaments, but at intervals during this treatment the heating was stopped and the tubes were allowed to cool to room temperature, fresh specimens of gas were admitted, and the Paschen curves then taken. After 116 hours of this procedure, the pure metal surfaces, viz. Ag in tube 1, Ni in tube 2, were subjected at intervals to ion bombardment, and the effect observed by taking Paschen curves. Since the bombardment by hydrogen ions was likely to set up a layer either of hydrogen atoms or of positive ions on the surface, the tubes were baked and the electrodes were strongly heated *in vacuo* for about 12 to 20 hours after each bombardment to ensure the removal of any dissolved hydrogen before the sparking potentials were measured.

This prolonged ion treatment of selected cathodes also had the effect of sputtering the metal on the opposite electrode (Llewellyn Jones and Davies 1951). In this way silver was deposited on the molybdenum electrode in tube 1 by making the silver the cathode in the hydrogen glow discharge; in tube 2 nickel was deposited on the aluminium electrode, and in tube 3 copper was deposited on the nickel electrode. In the present work it was found that a convenient rate of sputtering with copper and silver cathodes was obtained with glow discharge currents of 20 ma.; with nickel, however, the rate of sputtering was much lower, and currents of 100 ma. had finally to be used to produce observable deposits on the anode. This is in accordance with the work of Güntherschulze (1926), who gave the sputtering rate of nickel as approximately one-fifth that of copper or silver under similar conditions. Since the glow discharge between the electrodes was observed always to fill the space between the electrodes, it is reasonable to assume that the sputtering of the cathode occurred uniformly and that the metallic film was uniformly deposited on the anode. When the deposited films became thick enough to be visible, they did, in fact, appear quite uniform. Further, all these films were treated by mild hydrogen bombardment and outgassed with the filaments glowing. As with the pure metal electrodes, the Paschen curves in pure hydrogen were taken at various intervals during the formation and outgassing of the metallic film deposits.

§ 3. RESULTS

At these low pressures the breakdown relation for a given electrode separation d is $(\omega/\alpha)(\exp \alpha d - 1) = 1$, and since V_s and p were measured, X/p was known; with Hale's (1939) values of $\alpha/p = f(X/p)$, ω/α was found, and the values plotted as $(\omega/\alpha, X/p)$ curves.

The first significant result obtained was that the curves for copper, silver, and nickel in their initial states exhibited a pronounced peak; a typical example is given in Figure 1 for copper. These peaks are termed photoelectric peaks,

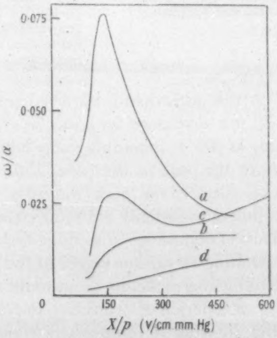


Figure 1. Effect of surface cleaning on secondary emission (ω/α) of copper: curve a, slightly oxidized; curve b, after baking; curve c, after 10 hours exposure to hydrogen; curve d, cleaned by ion bombardment and outgassed.

since a fall in ω/α with increase of X/p must be attributed to the photoelectric effects of the radiation emitted by gas atoms excited during the growth of the pre-breakdown ionization currents. On the other hand, a small steady increase of ω/α with X/p is generally accepted as an indication of the predominating influence of the impact of positive ions (Llewellyn Jones 1939).

The second significant result from these curves is that the removal of films from the cathode and subsequent outgassing removed the peaks. This effect was easily found with the copper surface but not so readily with silver and nickel surfaces. A copper surface covered with a thin oxide layer due to exposure to air showed a pronounced peak as indicated by curve *a* of Figure 1. After subsequent heating at 400° c. *in vacuo* for some 116 hours, curve *b* was obtained which exhibited no trace of a peak. Mild bombardment of the surface by hydrogen positive ions lowered the curve still further and finally altered its shape to one which showed a steady increase of ω/α with X/p throughout the range investigated (curve *d*).

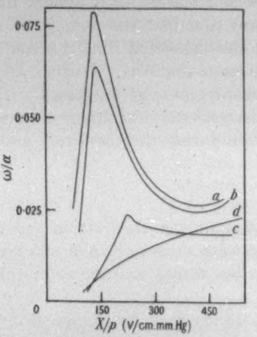


Figure 2. The effect on ω/α of deposition of a copper film on nickel: curve *a*, thin copper film; curve *b*, film outgassed; curves *c* and *d*, increase of film thickness and outgassing.

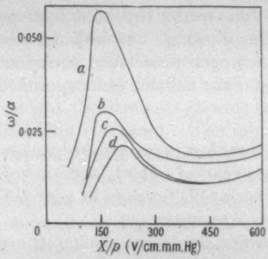


Figure 3. The effect on ω/α of the deposition of a silver film on molybdenum: curves *a* and *b*, thin silver film baked at 400° c. *in vacuo* for 48 hours and 116 hours respectively; curves *c* and *d*, further deposition of silver for 3½ and 6½ hours respectively and followed by outgassing.

The silver surface, on the other hand, required several periods of positive ion bombardment in a 20 ma. discharge for some hours followed each time by some 12 to 20 hours baking at 400° c. including many hours at red heat (to remove dissolved hydrogen) before the peak in the (ω/α, X/p) curve was considerably reduced. With the nickel surface, the peak was very difficult to remove, but continued severe bombardment eventually did reduce the peak considerably.

Consider now the effects of deposition of metallic films on the cathodes. The (ω/α, X/p) curves corresponding to various stages of the deposition of copper on nickel, silver on molybdenum, and nickel on aluminium are given in Figures 2, 3 and 4 respectively. After each period of deposition the electrodes were outgassed with the filaments glowing, in order to remove any dissolved hydrogen.

Figure 2 shows that when it was expected that the copper film was very thin (at the beginning of the deposition process), a sharp and pronounced peak in the secondary emission curve occurred. Curve *a* of Figure 2 gives the values of ω/α after the film and cathode has been baked at 400° c. for 116 hours *in vacuo.* Maintenance of the electrode *in vacuo* for a week had the result of lowering the values of ω/α to those of curve *b.* A further deposition of copper reduced ω/α to curve *c,* in which the peak was practically eliminated. Further deposition removed in one hour all traces of a peak (curve *d*), to give a curve which steadily increased with X/p and which was very similar to that obtained for a pure (bombarded) copper electrode illustrated in Figure 1.

It should be noticed that during the changes on the electrode surface (due to change in the film) the value of ω/α at the peak ($X/p \simeq 150$) was reduced by more than 90%, while at the higher values of X/p ($\simeq 400$) the values of ω/α were only reduced by about 20%. This result will be discussed in § 5.

With a silver film (Figure 3) the same general results were obtained, but the case of the nickel film (Figure 4) was different. Deposition of nickel for three hours had the effect of only reducing the peak by 25% (curve *b,* Figure 4) but deposition for another two hours reduced the peak by 70% (curve *c*): and further deposition altered the shape of the curve (curve *d*) by moving the peak to lower values of X/p and at the same time raising its value to 0·05, which was approximately the same as that for pure nickel. This final change in the electrode surface had a pronounced effect on V_m (390 v.), and it also reduced very markedly the value of ω/α at high values of X/p of the order of 300. The value of V_m for an aluminium cathode covered with a thin nickel film was about 390 v. in agreement with that found in the previous paper with different specimens of nickel and aluminium.

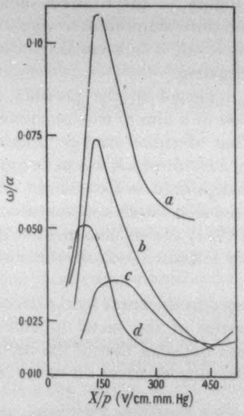


Figure 4. The effect on ω/α of deposition of nickel on aluminium: curve *a*, clean aluminium; curves *b*, *c* and *d*, increasing deposition of nickel for 3, 5 and 7 hours respectively and followed in each case by outgassing.

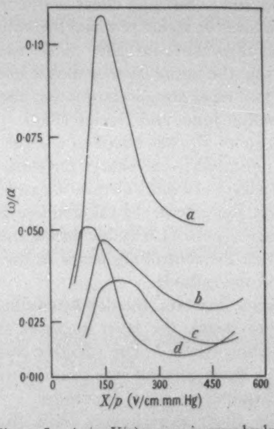


Figure 5. (ω/α, X/p) curves in pure hydrogen for cathodes of *a*, aluminium; *b*, nickel; *c*, thin nickel film on aluminium; and *d*, thicker nickel film on aluminium.

§4. THEORETICAL CONSIDERATIONS

The above experimental results are consistent with the view that secondary electron emission is greatly affected by the presence of an electrical layer, which, with its image, greatly reduces the work function (de Boer 1935). This is best illustrated by consideration of the cases of hydrogen and of thin metallic films on cathodes.

It is unlikely that positive ions or atoms of hydrogen remained on the cathode surfaces after the severe outgassing, and the lowest steady values of ω/α then obtained can be accepted as characteristic of emission from the pure surface. Prolonged exposure to hydrogen will have an effect on ω/α which depends upon the electrical polarity of the adsorbed layer. Hydrogen atoms will tend to lose an electron to the more electro-negative copper or silver to leave a positive layer, and the secondary emission will consequently be increased. This is the result shown in Figure 1, curve *c*.

On the other hand, if a monatomic layer of nickel atoms acquire electrons from the underlying aluminium to set up a negative layer, the emission would be reduced. Such an effect is illustrated in Figure 5: these curves show how a thin layer of nickel reduced the emission from pure aluminium to values lower in fact than those for pure nickel itself. This result is in accordance with the view that the initial layer of nickel atoms was negative.

This view also accounts for the results reported in the previous paper (Llewellyn Jones and Davies 1951). In the case of a film of iron on aluminium the value of V_m was raised to a value above that of either steel or aluminium. This indicated an increase in the work function, a result which was to be expected when a layer of a more electro-negative iron was deposited on aluminium. Again, the high increase in the value of ω/α as X/p is reduced for air contaminated with mercury vapour (Llewellyn Jones and Davies 1951) is also consistent with the view that the controlling factor in the secondary ionization was the thin mercury film on the cathode.

When, however, the deposited film is increased in thickness by the deposition of more metal, the layer acquires the properties of the metal in bulk with conduction bands. The effective work function becomes that of the deposited metal. Such a result was found with thick deposits of silver on molybdenum and of copper on nickel electrodes.

The above experiments show that the state of the cathode surface plays a far greater part in the mechanism of the low pressure discharge than had perhaps hitherto been realized. The removal of thin oxide films from, and the deposition of thin metallic films on, cathodes had a profound effect on the $(\omega/\alpha, X/p)$ curves. It is considered that the deposition of a thin film on a more electro-positive metal as cathode raised the work function, while a thin film on a more electro-negative metal lowered the work function. Each of these changes in the surface work function were accompanied by a very rapid change in the magnitude of the photoelectric peaks in the $(\omega/\alpha, X/p)$ curves ($X/p \simeq 150$ v/cm.mm. Hg) while little change occurred in that part of the curves beyond $X/p \simeq 400$ v/cm.mm. Hg. When the cathode surface was considered to be in a pure state, the photoelectric peak was absent, and the $(\omega/\alpha, X/p)$ curves increased steadily with X/p throughout the range. This form of curve is generally explained by attributing the secondary emission to the impact of ions, the energies of which we expect to increase with X/p.

§ 5. THE SECONDARY IONIZATION PROCESS

Information concerning the relative importance of the two secondary ionization processes (δ and γ) in spark discharges can be obtained from an interpretation on the following lines of the observed effects of surface films on the cathode electron emission.

It can be shown that the photoelectric current i_{ph} from a metal due to the incidence of black-body radiation at a temperature T can be represented by a relation (de Boer 1935)

$$(i_{\mathrm{ph}})_0 = AT^2 \exp(-\boldsymbol{e}\phi_0/\boldsymbol{k}T), \qquad \ldots\ldots(1)$$

where A is a constant, and ϕ_0 the work function of the metal (3 to 4·5 ev. for common metals), and $\boldsymbol{k}$ is Boltzmann's constant. During the sputtering of the anode and the deposition of atoms (or ions) of that metal on the cathode surface it can be assumed that the effective work function of the whole cathode surface was being gradually changed. It follows from the work of de Boer and Bruining (1939) that it may be assumed that

$$\phi_t = \phi_0 - ct, \qquad \ldots\ldots(2)$$

where c is a constant which is positive or negative depending on the polarity of the electrical double layer formed on the cathode surface. Consequently, the corresponding photoelectric current at time t is given by

$$(i_{\mathrm{ph}})_t = AT^2 \exp\{-\boldsymbol{e}(\phi_0 - ct)/\boldsymbol{k}T\} = (i_{\mathrm{ph}})_0 \exp(\boldsymbol{e}ct/\boldsymbol{k}T). \qquad \ldots\ldots(3)$$

From this it is seen that if $\boldsymbol{k}T < \boldsymbol{e}ct$ the influence of the exponential factor can be very great, and the photoelectric current would then change rapidly with time during the deposition of the surface layer. This change would continue until the surface layer became thick enough to form conduction bands and the work function of the surface then characteristic of the deposited metal in bulk. On the other hand, if $\boldsymbol{k}T$ were large compared with $\boldsymbol{e}ct$ then equation (3) would become (neglecting second order terms)

$$(i_{\mathrm{ph}})_t = (i_{\mathrm{ph}})_0(1 + \boldsymbol{e}ct/\boldsymbol{k}T), \qquad \ldots\ldots(4)$$

and from this equation it can be seen that the change in the photoelectric current with deposition would be relatively small when $\boldsymbol{k}T \gg \boldsymbol{e}ct$.

Thus the energy of the incident photons determines the rate of variation of the secondary emission with change of surface work function. With low-energy photons there is a large change of δ with ϕ, while with high-energy photons the change of δ with ϕ is small.

Now the total radiation from hydrogen atoms and molecules produced during the growth of the pre-breakdown ionization currents is not the same as that from a black body at a temperature corresponding to the average energy of all the emitted photons. However, there are a very large number of photons of low energy due to transitions to the first resonance band, and for the purpose of these general considerations no great error is made by regarding the total radiation in this low-energy group of photons as being similar to that from a black body at the temperature T given by

$$\boldsymbol{k}T = 2\boldsymbol{e}E_1/3, \qquad \ldots\ldots(5)$$

where $\boldsymbol{e}E_1$ is the mean energy of the low-energy group of photons ($\sim$1·5 v.). To this group, therefore, equations (3) and (5) will be considered as approximately applicable, giving

$$(i_{\mathrm{ph}})_t = (i_{\mathrm{ph}})_0 \exp(3ct/2E_1). \qquad \ldots\ldots(6)$$

If now the rate of deposition of the surface film is such that when $E_1 \simeq 1{\cdot}5$ v. $ct \geqslant E_1$, it follows from (6) that electron emission due to the impact of low-energy photons increases rapidly with the formation of a positive layer (c positive) and similarly decreases rapidly with the formation of a negative layer (c negative).

For the high-energy photons, on the other hand, produced by transitions to the ground state, these can be regarded as forming a group of mean energy $\boldsymbol{e}E_2 \simeq 12{\cdot}5$ ev., and equation (4) then gives

$$(i_{\mathrm{ph}})_t = (i_{\mathrm{ph}})_0(1 + 3ct/2E_2). \qquad \ldots\ldots(7)$$

Since E_2 is now likely to be much greater than ct, the change of photo-emission with deposition of a layer is then small when that photo-emission is due to the high-energy photons.

Consider now secondary emission due to the impact of ions of high potential energy (of the order of 15 ev.). It has been shown by Sixtus (1929) that the secondary emission i_γ due to impact of electrons can be given by an equation of the form

$$(i_\gamma)_0 = P \exp(-\boldsymbol{e}\phi_0/\boldsymbol{e}E), \qquad \ldots\ldots(8)$$

where P is a constant and $\boldsymbol{e}E$ the mean energy of the emitted electrons. The effect of the deposition of a surface film on such emission is found by using equations (2) and (8), from which it follows that the change of emission with time is given by the relation (de Boer and Bruining 1939)

$$(i_\gamma)_t = (i_\gamma)_0 \exp(\boldsymbol{e}ct/\boldsymbol{e}E), \quad \text{or} \quad (i_\gamma)_t = (i_\gamma)_0(1 + ct/E), \qquad \ldots\ldots(9)$$

when $E > ct$. This relation, as for (4), indicates that the change of secondary emission with time is linear and relatively small. It will here be assumed, as an approximation, that a relationship such as (8) holds also for secondary emission due to the impact of ions, when the emitted electrons have high energies ($E \sim 6$ v.) (Oliphant 1930); it is therefore interesting to apply this result to the γ process.

From all the above considerations it then follows that a slow linear change of ω/α with time during the formation of a surface film is consistent with the view that secondary emission is due to the impact of ions or of high-energy photons; on the other hand, a rapid change of emission with time is consistent with the view that ω/α is primarily due to the impact of low energy photons (visible range).

Consider now the experimental results:

Just as the deposition of positive ions on a surface should rapidly increase low energy photo-emission, so the gradual elimination of the surface layer should produce a rapid diminution in the photoelectric peak. Figure 1 shows how the photoelectric peak did, in fact, change with time during the outgassing and surface cleaning; and it can be seen that the photoelectric peak was almost completely eliminated, while at the same time that part of the emission ω/α which could definitely be ascribed to secondary emission due to the impact of ions ($X/p \sim 400$ v/cm.mm.Hg) did not change greatly. This result is clearly seen from Figure 6 which shows the variation of ω/α (observed at values of X/p of 150 and 400 v/cm.mm.Hg) with the cleaning time for these three cathodes. The conclusion follows that at gas pressures in the neighbourhood of the minimum sparking potential the photons responsible for the photoelectric peak were of low energy (visible range): thus $\delta > \delta'$ or $\alpha\gamma$. It follows that if any high-energy photons were produced during the ionization leading to breakdown, either they were rapidly degraded in transmission through the gas, or the probability of their production was small. As, however, the gas pressure was low ($\simeq 5$ mm.Hg) and the distance between the electrodes small ($=0{\cdot}20$ cm.) degradation of photons would not be great and it may therefore be concluded that the production of them was small. It also follows from the above considerations that the predominant secondary mechanism with the metals investigated in the pure state (high ϕ) was cathode emission due to the incidence of positive ions or of excited atoms; since, however, all such cases showed a steady increase of ω/α with X/p, this increase supports the view that the impact of positive ions was more important than that of excited atoms ($\alpha\gamma > \delta'$).

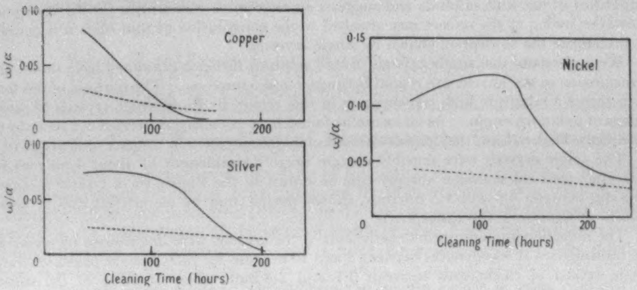


Figure 6. The variation of ω/α measured at $X/p \simeq 150$ (full curve), and at $X/p \simeq 400$ (dotted curve), with the time of cleaning of the cathode surface, i.e. removal of any surface films.

It is concluded that the mechanism of low-pressure breakdown in hydrogen (and probably also in other similar gases) is practically controlled by the surface properties of the cathode. Low-energy and high-energy ionizing photons, positive ions, and excited atoms are all produced during the growth of pre-breakdown ionization currents in the gas, but the role they actually play in causing breakdown is mainly determined by the cathode surface. This result is relevant to consideration of the mechanism of Geiger counters.

ACKNOWLEDGMENTS

We wish to thank the Department of Scientific and Industrial Research for the award of a grant to one of us (D. E. D.), and Messrs. H. Wiggin and Sons for the gift of pure nickel sheet.

REFERENCES

de Boer, J. H., 1935, *Electron Emission and Adsorption Phenomena* (Cambridge : University Press).

de Boer, J. H., and Bruining, H., 1939, *Physica*, **6**, 941.

Günterschulze, A., 1926, *Z. Phys.*, **36**, 563; **38**, 575.

Hale, D. H., 1939, *Phys. Rev.*, **55**, 815; **56**, 1199.

Llewellyn Jones, F., 1939, *Phil. Mag.*, **28**, 192, 328.

Llewellyn Jones, F., and Davies, D. E., 1951, *Proc. Phys. Soc.* B, **64**, 397.

Llewellyn Jones, F., and Henderson, J. P., 1939, *Phil. Mag.*, **28**, 185.

Oliphant, M. L. E., 1930, *Proc. Roy. Soc.*, **127**, 373.

Sixtus, K., 1929, *Ann. Phys., Lpz.*, **3**, 1017.

LETTERS TO THE EDITOR

Infra-Red Transmission of Galena

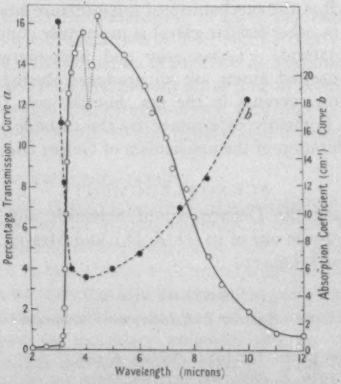
The considerable interest in the photoconductivity of lead sulphide in recent years has made the investigation and interpretation of its absorption spectrum important.

Gibson (1950) has described work on the absorption spectra at different temperatures of thin films of lead sulphide, lead selenide and lead telluride. According to Gibson the lattice absorption edge for lead sulphide films falls at a wavelength of about 0·9 micron and is followed by a tail absorption band reaching to 6 microns wavelength with an absorption coefficient of the order of 10^4 cm^{-1}. He considers that this band is unlikely to be due to impurities in the lead sulphide and suggests an alternative mechanism for the absorption. However, owing to the uncertainty attached to the constitution of thin films it is desirable to investigate the absorption shown by single crystals.

We understand that single crystals of lead sulphide hitherto examined have shown little transmission in the wavelength region between 3 and 6 microns. The purpose of this letter is to report a relatively high transmission in this region by some single crystals of natural galena of unknown origin. As far as can be found by x-ray analysis, carried out for us by the Chemistry Department, they possess no detectable impurity.

The single crystals were sensibly opaque up to a wavelength of about 3 microns but thereafter their transmission sharply rose as shown in the Figure for a typical specimen. The dip between 4·0 and 4·5 microns, shown dotted, may be an artifact due to carbon dioxide absorption in the atmosphere.

The reflection and absorption coefficients of the crystals were determined by measuring the transmission at wavelengths between 3 and 10 microns for each of eight pieces of cleaved single crystal of thicknesses between 0·5 and 2·8 mm. The variation of the derived absorption coefficient with wavelength is shown in the Figure. The reflection coefficient was constant within 10% at 0·4, corresponding to a refractive index of 4·4. The values of both coefficients are accurate only to within 10%.



The rise after 3 microns may well be significant, since it coincides with the sharp fall in the photoconductivity of lead sulphide at room temperature shown by thin films, and in the photovoltaic effect shown by crystals using point contacts (Fischer, Gudden and Treu 1938).

Polished samples from another source showed little transmission. Their opacity was not due to the fact that the surfaces were polished instead of being cleaved, since a cleaved specimen of the first material showed no change in transmission after polishing. The sub-crystal structure (small-angle grains) was very much more marked in the second type of galena, which is consistent with either rapid growth or the presence of impurity. The absorption might be due directly to the impurity or perhaps to some process at the grain boundaries. The amount of impurity was too small for detection by x-ray analysis.

The transparent specimens of lead sulphide are similar to silicon and germanium (Becker and Fan 1949) both in the shape of the absorption curve and in the fact that absorption falls away sharply above the photoconductive limit. If the resemblance extends to lead selenide and lead telluride, these latter materials might be expected to transmit at about 5 microns. It is of interest that a single crystal of lead selenide given to us by Dr. R. A. Smith of the Telecommunications Research Establishment has shown the expected rise in transmission at about 5·2 microns.

Natural Philosophy Department,
University of Aberdeen.
10th March 1951.

W. PAUL.
D. A. JONES.
R. V. JONES.

BECKER, M., and FAN, H. Y., 1949, *Phys. Rev.*, **76**, 1530.
FISCHER, F., GUDDEN, B., and TREU, M., 1938, *Phys. Z.*, **39**, 127.
GIBSON, A. F., 1950, *Proc. Phys. Soc.* B, **63**, 756.

Magnetic-Tape Recording of Electrostatic Field Changes

Two methods of photographic recording of the output of an electrostatic fluxmeter, for the measurement of electric fields produced by thunderstorms, have recently been described by Malan and Schonland (1950). In the first, an anticipatory lightning trigger, operating in the initial stages of the stepped leader process, biases the oscilloscope and automatically shifts the trace from behind a mask for a predetermined time interval. Hence, field-changes immediately before and during the earliest stages of the stepped leader process are not recorded. In addition the trigger circuit is often operated by distant flashes to whose field-changes the fluxmeter is insensitive. In the second method the gain of the oscilloscope is manually and continuously adjusted to keep the peak of the trace behind a small mask at the centre of a slit, so that it only emerges for rapid field-changes. Small, unwanted field-changes are often unavoidably recorded and unless the drum camera is traversed axially after each such recording these cause annoying overlapping.

Both methods are extravagant as regards the amount of film used, and the method of recording now described aims at economy in this respect and the elimination of the other disadvantages mentioned above.

The 1,200 c/s. output from the fluxmeter is first fed through a pre-amplifier and then in parallel to a commercial type magnetic-tape recorder and an oscilloscope. By means of the latter, the output of the fluxmeter may be checked visually.

In this way a continuous recording of field intensity is obtained, it being necessary to reload the recorder once every half an hour only, an operation taking a few minutes. Coding of the field-changes is effected by switching a microphone into the recording circuit and at the same time removing the fluxmeter output. A verbal description of the flash and its distance thus follows each field-change.

Permanent records are made by playing back selected portions of the tape into an oscilloscope and photographing the screen by means of a revolving drum camera. The camera revolves once every four seconds, and by suitably marking the tape and using a manually operated shutter, it is possible to record field-changes from about one second before, to two or three seconds after, the main field-changes.

Because of the comparatively low output from the fluxmeter it has been found necessary to ensure the smooth running of the tape and to pay particular attention to the elimination of unwanted ripple throughout the apparatus.

The Bernard Price Institute of Geophysical Research,
University of the Witwatersrand,
Johannesburg.
21st February 1951.

N. D. CLARENCE.
D. J. MALAN.

MALAN, D. J., and SCHONLAND, B. F. J., 1950, *Proc. Phys. Soc.* B, **63**, 402.

Poisson's Ratio and Electrical Resistance

If a wire is subjected to a stress its linear dimension and its electrical resistance change, and it should be possible to make use of the latter alteration to determine Poisson's ratio for the material.

Consider an isotropic wire of length a, diameter d, electrical resistance r and specific resistance s. We shall denote changes in these quantities by the suffix 1 for axial stress, and suffix 3 for bulk stress. Let Poisson's ratio be σ. Then, since $r \propto sa/d^2$,

$$\delta(\log r)_1 = \delta(\log s)_1 + (1+2\sigma)\delta(\log a)_1. \qquad \ldots\ldots (1)$$

The quantity $\delta(\log r)_1/\delta(\log a)_1$ is easily measured (by making a Young's modulus experiment with the specimen forming one arm of a Wheatstone bridge), but while Bridgman has measured $\delta(\log s)_3/\delta(\log a)_3$ for most metals and a number of alloys, there are few values of $\delta(\log s)_1/\delta(\log a)_1$ available. It should be possible to deduce σ however from the Bridgman data and a measure of the change in resistance of the wire under pressure : for, in the case of bulk stress, $\delta(\log r)_3 = \delta(\log s)_3 - \delta(\log a)_3$ and $\delta(\log r)_3/\delta(\log a)_1 = (1-2\sigma)(3B-1)$, where $B = \delta(\log s)_3/\delta(\log v)$ and $\delta(\log v)[=3(1-2\sigma)\delta(\log a)_1]$ is the fractional change in volume under bulk stress : B may be termed the Bridgman coefficient.

Difficulty arises in practice because $\delta(\log r)_3$ is very small, and this method is only suitable where facilities are available for the application of pressures of, say, 1,000 atmospheres.

An alternative procedure is possible if we make the assumption that the effective change in cross section which an atom presents to the passage of an electron is independent of the way in which the stress producing the change is applied. Assuming that the specific resistance of the metal is some function, ϕ say, of the cross section of the atoms, and varies as the number of atoms per unit volume, i.e. for a given specimen $s \propto \phi/ad^2$,

$$\delta(\log s)_1 = \delta(\log \phi)_1 - \delta(\log a)_1 + 2\sigma\delta(\log a)_1$$

and $\delta(\log s)_3 = \delta(\log \phi)_3 - 3\delta(\log a)_3 = \delta(\log \phi)_3 - 3(1-2\sigma)\delta(\log a)_1$ and, therefore, remembering that $\delta(\log \phi)_1 = \delta(\log \phi)_3$,

$$\delta(\log s)_1 - \delta(\log s)_3 = 2(1-2\sigma)\delta(\log a)_1 \qquad \ldots\ldots (2)$$

Using (2) in (1), we find that $\delta(\log r)_1/\delta(\log a)_1 = 3 - 2\sigma + 3(1-2\sigma)\delta(\log s)_3/\delta(\log v)$. Writing R for the strain sensitivity, $\delta(\log r)_1/\delta(\log a)_1$, we then obtain

$$\sigma = \frac{3(B+1)-R}{2(3B+1)}. \qquad \ldots\ldots (3)$$

With the exception of gold, which was not annealed, R has been directly measured before and after annealing for five wires for which published values of B were available (Mott and Jones 1936), and the values of σ calculated from equation (3) are shown in the Table:

		Before annealing			After annealing		
	B	R	σ	σ'	R	σ	σ'
Cu	2·7	2·7	0·46	0·08	3·2	0·43	0·41
Ni	3·3	4·3	0·40	0·18	6·2	0·31	0·28
Ag	3·6	3·7	0·43	0·56	3·2	0·45	0·36
Au	5·2	6·8	0·36	0·31	—	—	—
Pt	5·6	4·6	0·43	0·59	4·9	0·42	0·33

values of σ obtained with the same wires from measurements of Young's modulus and the coefficient of rigidity, using a dynamical method for the latter, are also shown under the columns headed σ'. The wires were annealed with a hand torch until no further change took place in the measured values of the moduli. The nickel, silver, gold and platinum wires used were of fairly high purity and the copper was cut from a reel of commercial cotton-covered wire.

The results for silver and copper agree reasonably well with those obtained by Linde, von Heijne and von Sabsay (1950), but the value for gold obtained by Linde and Stade (1950) differs widely from ours.

We should like to express out thanks to the Director of the South African Mint, Pretoria, for supplying us with pure gold wire, and to Mr. W. R. McMurray and Mr. P. Weedon for help withthe experiments.

Department of Physics,
University of Natal,
Pietermaritzburg,
Natal, S. Africa.
16th April 1951.

J. R. H. Coutts.*
J. A. V. Fairbrother.

Linde, J. O., von Heijne, A., and von Sabsay, E., 1950, *Ark. Fys.*, **2**, 81.
Linde, J. O., and Stade, C. H., 1950, *Ark. Fys.*, **2**, 99.
Mott, N. F., and Jones, H., 1936, *Properties of Metals and Alloys* (Oxford : University Press), p. 272.

* Now at Marischal College, University of Aberdeen.

Discussion

on papers by J. B. Higham and J. M. Meek entitled "Voltage Gradients in Long Gaseous Spark Channels" and "The Expansion of Gaseous Spark Channels" (*Proc. Phys. Soc.* B, 1950, **63**, 633 and 649); *read at Science Meeting of the Society on 24th November* 1950.

Dr. A. von Engel (*Clarendon Laboratory, Oxford*). In hydrogen at 1 atm. the electric field in the spark channel changed from 10^5 v/cm. to less than 10^2 v/cm. in $\Delta t \leqslant 0{\cdot}2\,\mu$sec. With an electron drift velocity of order 10^7 cm/sec. I find that the electron transit time for 10 cm. electrode distance is about the same value.

After $2\,\mu$sec. the electric field in the gas becomes constant (about 35 v/cm.) and so does the channel diameter. Since the field in a steady arc column—enclosed in a tube of 4 cm. diameter and stabilized by gas currents—is about 50 v/cm. at the same current, I conclude that the spark channel in H_2 for $\Delta t > 2\,\mu$sec. can be regarded as a positive column.

In air the spark is still developing after $20\,\mu$sec. and then shows 50 v/cm. as compared with a few v/cm. for the steady arc column. The difference in the 'time constants' seems to be significant. If the instantaneous power could be measured, this, together with the energy balance, would give further information about the main ionization processes and losses.

Diffusion cannot explain why the channel radius becomes constant so quickly. I suggest that a shock wave travelling with a speed above that of sound would produce such a rapid expansion.

Professor F. Llewellyn Jones (*Department of Physics, University College of Swansea*). The experiments described in these two papers provide interesting electrical data concerning the characteristics of sparks at the higher pressures. However, the fact that the applied E.M.F. was impulsive does introduce some difficulty of interpretation : also the spark length apparently was not determined by the distance between the electrodes, and this effect introduced some uncertainty in the value of the electric field Z deduced. The experiments nevertheless appear to establish empirical relationships between the current i, the average electrical field Z, and the gas pressure p. Z (except for hydrogen) is independent of i, and i for any spark is proportional to the area of cross section A : but it is not easy to understand the apparent independence of Z and p. Is it possible that any variation in Z was too small to be measured under the conditions of the experiments ?

It is interesting to consider whether these data can provide us with information concerning the discharge processes in this high pressure range. The fact that Z is independent of i supports the conclusion that recombination and other processes which are not proportional to the current are not important compared with those processes, such as single impact ionization by electrons or diffusion, which are proportional to the current. Consequently, it follows that lateral diffusion plays an important part in the discharge mechanism, and this conclusion is supported by the dependence of Z on the cross section A

when this is artificially restricted. In view of these facts it is interesting to consider whether any similarity relationship might hold for the discharges investigated. Such relationships are the following :

$$\phi(J/p^2)=\text{constant} \quad \ldots\ldots(1) ; \quad \text{and} \quad p\sqrt{A}=f(Z/p). \quad \ldots\ldots(2)$$

Consider the first relationship. Since $i=AJ$, and if (J/p^2) is constant, then $i/Ap^2=\text{constant}$. If i only increases slowly with p, it would follow that A should diminish as p increases. It is interesting to note that one of the observed results was a noticeable decrease of A as p increased. Z here of course means the actual field in the glow, but the Z measured was V/d, where d is the computed spark length and V the sparking potential ; then if the authors in continuing this work find they can measure d and Z more precisely, it would be interesting to test the relationships of (1) and (2) more specifically.

I agree with the suggestion that the channel cross section may not be the same as the discharge diameter. Experiments performed in Swansea on high-frequency corona breakdown at the lower pressures showed clearly that the discharge occupied a volume much greater than the luminous region, and that lateral diffusion from a glowing column was most important in the discharge mechanism.

Dr. J. D. Craggs (*Department of Electrical Engineering, University of Liverpool*). There are many aspects of the work on which I should like to comment, but I shall restrict my remarks to certain matters relating to the physical properties of spark channels in hydrogen at currents of about 100 amp. The following has a direct bearing on Mr. Craig's remarks.

The work by Mr. Hopwood and myself on the measurement of ion or electron concentrations n_e by Stark effect observations has been extended theoretically by Dr. Grimley and Dr. Saxe and experimentally by ourselves and the latest results suggest, for a particular form of spark, that the mean n_e is close to $10^{17}/\text{cm}^3$ and the corresponding electron drift velocity v_d is about 7×10^5 cm/sec. at 2 μsec. after breakdown.

Next, from a knowledge of the voltage gradients in the same discharge, the electron drift velocities may be calculated. We have assumed that electron energy loss by inelastic collisions is negligible and we have computed the Gvosdover scattering due to the positive ions. The drift velocity for a constant pressure channel agrees well with the 'Stark v_d' for a channel temperature ~15,000° K. but is ~1% of the 'Stark v_d' at the same temperature in a constant density channel (Gvosdover scattering almost negligible). In the latter case it would appear that an unreasonably high temperature would be required to give agreement, and so indirect support for a constant pressure regime at 2 μsec. is provided.

Mr. Craig and I hope to publish an account of this work in due course.

Mr. R. D. Craig (*Department of Electrical Engineering, University of Liverpool*). Referring to the oscillographic measurements of the voltage across spark channels in glass tubes an alternative, and possibly more probable, explanation of the wide difference between the diameters as deduced from these experiments and those later obtained for the corresponding luminous channel in free air is that in the former experiments the observed increase in voltage may have been caused by the passage of a shock wave, generated in the initial stage of the discharge and subsequently reflected back from the walls of the tube. It is of interest to note that Foley (1920) photographed the shock waves arising from the initiation of a spark discharge in air and that he estimated the mean velocity over the first few millimetres to be of the order of 10^5 cm/sec. as would be required if the above explanation applied to the authors' experiments; also, Suits (1936) has shown that the effect of the passage of a shock wave through an arc discharge is to cause an increase in the burning voltage. It may also be pertinent to observe that Mason (1937) has shown by probe experiments that in the case of the continuous arc there is no appreciable difference between the effective electrical boundary and that of the luminous core.

With regard to the paper on the expansion of spark channels recent work carried out in Liverpool in conjunction with Dr. J. D. Craggs indicates that of the various possible expansion mechanisms suggested by the authors the predominant one during the initial stages appears to be that of mass flow arising from pressure inequalities produced by the rapid heating of the gases in the discharge channel. It would seem therefore that conditions during the first μsec. or so may best be treated by shock wave theory. An important consequence of such a treatment is that approximate pressure equilibrium and hence arc-like conditions of relatively low particle concentration are probably reached within a matter of microseconds rather than within times of the order of a millisecond as has often been assumed.

Authors' Reply. The lack of an adequate theoretical treatment of the mechanism of expansion of a spark channel has undoubtedly handicapped the interpretation of our results, and we have therefore been interested to have the suggestions made by the several contributors to the discussion. It is clear that certain differences of opinion exist. While diffusion is stated to be important by Professor Llewellyn Jones this effect is discounted by Dr. von Engel who favours the view held by Mr. Craig, that the initial expansion of the spark channel may be explained by consideration of shock wave theory. We look forward to the publication of further details of the analysis of spark channels on the latter basis.

The need for more exact measurements of the spark length and the voltage drop has been mentioned by Professor Llewellyn Jones. However, we doubt whether improvement in accuracy would be greatly advantageous because of other uncertainties: for example, we do not know the variation of voltage gradient along the spark path, in which many bends occur.

The results quoted are the average of many measurements and we should expect changes caused by variations of a parameter such as pressure to be apparent if they amount to more than about 5%. Within the pressure range investigated we were unable to establish any trends in the relation between the voltage gradient and the gas pressure. For pressures between $\frac{1}{2}$ and 1 atmosphere there is no change in the photographed channel diameter for a constant current, and therefore the similarity relationship i/Ap^2=constant does not appear to hold in this pressure range.

Below 100 mm. Hg the discharges have a different character and, as mentioned in the papers, the present techniques are unsuitable. The measurements will be extended to pressures above 1 atmosphere in order to investigate more adequately the dependence on pressure. In future work it is probable that some results will be obtained in which each spark is photographed from two directions at right angles at the same time as the voltage drop is being measured, so that direct correlation can be obtained between the values for the voltage drop and the spark length for individual sparks rather then the mean values for a group of sparks.

Foley, A. L., 1902, *Phys. Rev.*, **16**, 449.
Mason, R. C., 1937, *Phys. Rev.*, **51**, 28.
Suits, C. G., 1936, *Gen. Elect. Rev.*, **39**, 430.

REVIEWS OF BOOKS

Techniques générales du laboratoire de physique, Volume II, edited by J. Surugue, Pp. 336. 1st Edition. (Paris: Centre National de la Recherche Scientifique. 1950; obtainable in London through Messrs. H. K. Lewis.) In paper 42*s*.; in cloth 44*s*.

The second volume of *Techniques générales du laboratoire de physique*, edited by Professor J. Surugue, contains sections on the measurement of radiant energy, on quartz fibres and Wollaston wires, thin cellulose films, electrometers, electrometer valves, ionization counters, the technique of electrical analogies, micromanipulation, and automatic temperature control. Each section is written by a separate author, and there is much individuality of treatment. The result is not a coordinated handbook, but a sequence of monographs on subjects of widely differing weight, where in each case the author is writing about a field in which he has had direct experience.

Approximately 100 pages (about one-third of the volume) are devoted to the section on measurement of radiant energy, by J. Lecomte. In this section much more attention is given to non-selective detectors such as thermopiles than to selective detectors of the photoelectric type, but the section is particularly valuable for drawing attention to many original papers which are often overlooked in more cursory surveys. A point of criticism, which applies to the whole volume, is that the references are given unobtrusively in the text and are not easy to find again quickly. A pleasing feature of this section is its mention of relatively unconventional and little-known devices for detecting radiant energy, such as Tyndall's 'singing flask', which was the forerunner of the modern pneumatic detectors. At the same time there is a fair selection of references to recent work.

In the appropriate sections an adequate treatment is given of techniques for measuring small currents both by deflection electrometers and also by electrometer valves. A useful review is made of the factors setting the ultimate limit of sensitivity of such valves and of balanced circuits for securing stability of zero. The theory and design of ionization chambers is not discussed.

The section on particle counters is of necessity summary, since the field is large and rapidly expanding. Even so, mention might have been made of the Rosenblum counter, of halogen-quenched low-voltage counters, and of cold cathode valves as useful adjuncts in counting circuits. It would enhance the value of this section, and of that on electrometer valves, to include tabulated data of the operating characteristics of the counting tubes and valves commercially available.

The section on micromanipulation by P. de Fonbrune is particularly worthy of note.

The volume suffers from a very inadequate index. Many instruments known universally by the name of their designers are described in the text, but few can be found in the index. The Lauritsen electroscope is one example of many.

Despite this drawback, Professor Surugue and his colleagues have produced a useful account of some important techniques in experimental physics, and have in several cases brought together a range of information hitherto not available in a compact form. The book will be surely welcomed in any laboratory that is likely to use these techniques; we look forward to subsequent volumes in the same series. R. V. JONES.

Electrophysiological Technique, by C. J. DICKINSON. Pp. vii + 141. (London: Electronic Engineering, 1950). 12*s.* 6*d.*

At first sight this laboratory manual would not be expected to have value outside the specified field of the title. The author is, however, too modest in this respect; almost the whole manual is equally valuable in a host of other fields. Indeed, in a sense it is calculated to give an almost alarming impression of the complexity and range of technique which an electrophysiologist should have at his finger-tips. Not all the luxuries described may be within the resources of potential entrants to the field, and much has been, and still can be, accomplished with far less. This, however, is hardly a criticism. If some of the functions described could be performed by utility circuits using fewer valves, the author has the defence that his circuits at least gain in predictability by their occasional lavishness (the crystal calibrator is a case in point).

Mr. Dickinson occasionally shows too uncritical a spirit. It is unfortunate, for example, that one of his two recommended power stabilizers has the defect of using the negative supply line as a standard for the positive stabilizer (Figure 5, p. 18). The practice of under-running valve heaters is recommended without comment. One is surprised to be told on page 66 that " It is generally unnecessary and complicated to introduce feedback into A.C. amplifiers ". Has this any connection with the fact that ' overall ' feedback was recommended in preference to stage-by-stage degeneration ? In fact it is not at all difficult to achieve the benefits of feedback in A.C. amplifiers, particularly in the reduction of microphony. (Incidentally where did the author find the term " microphonicity " ?)

The discussion of noise level in galvanometers on page 79 is practically meaningless in the absence of figures of sensitivity. It is all too easy to build a galvanometer whose " noise level is so small as to be almost unmeasurable " ! What are " thermal emission effects " in a galvanometer ? Should this be a reference to photocells ?

A number of misprints in the text and illustrations (e.g. screen supply in Figure 39, p. 92) will doubtless be corrected in a future edition, and it is to be hoped that occasional lapses into electronic slang will also be amended.

These are, however, somewhat minor criticisms. As a whole the book succeeds well in providing the much-needed facilities it sets out to offer. Its usefulness is much enhanced by the adoption of the admirable ' Electronic Engineering ' practice of providing circuit values with the diagrams given. The suggestions about valve standardization though somewhat arbitrary, will go far towards giving the newcomer to electronics the perspectiv he needs.

J. T. RANDALL

CONTENTS FOR SECTION A

PAGE

Dr. L. Kellner. The Vibrations of an Infinitely Long Chain of CH_2-Groups and the Infra-Red Spectrum of Polythene 521

Dr. C. Joan D. Jarvis and Dr. M. A. S. Ross. Modes of Disintegration of Ionium : an Investigation using β-sensitive Emulsions 535

Mr. W. F. Caplehorn and Mr. G. P. Rundle. The Angular Distribution of 3 MeV. Neutrons scattered by Protons and by Deuterons 546

Dr. G. K. White. Flow of Liquid Helium through Fine Channels . . . 554

Mr. R. Bowers and Dr. G. K. White. Pressure Gradients in Superflow . . 558

Dr. H. Edels and Dr. J. D. Craggs. Excitation Temperatures of Hydrogen Arcs 562

Dr. H. Edels and Dr. J. D. Craggs. Balmer Line Widths in Hydrogen Arcs . 574

Letters to the Editor :

Dr. J. H. E. Griffiths and Mr. J. Owen. Paramagnetic Resonance in Nickel Ammonium Sulphate 583

Mr. F. A. El-Bedewi. The Range of Protons from the Reaction ^{14}N (n, p) ^{14}C in Ilford C2 Photographic Emulsion 584

Mr. D. West and Dr. J. K. Dawson. Soft Radiations from ^{239}Pu . . . 586

Mr. J. E. Allen. The Distribution of Electron Energies in a Discharge constricted by its Self-Magnetic Field 587

Mr. A. Lempicki. The Electrical Conductivity of Simple p-Type Semiconductors 589

Mr. T. S. Moss. Photoconductivity in the Elements 590

Mr. W. R. S. Garton. Ultra-Violet Absorption Spectra of Tin Vapour in Atmospheres of Helium and Hydrogen 591

Dr. M. W. Feast. On a Band System ascribed to the CH Molecule . . 592

Reviews of Books 593

Contents for Section B 598

Abstracts for Section B 599

ABSTRACTS FOR SECTION A

The Vibrations of an Infinitely Long Chain of CH_2-Groups and the Infra-Red Spectrum of Polythene, by L. Kellner.

ABSTRACT. An attempt has been made to interpret the infra-red spectrum of polythene as the spectrum of an infinitely long chain of CH_2-groups. The vibrations of such a chain have been calculated using a potential function which involves valence and angle forces as well as the torsional force resisting a twisting of the C–C bonds. The selection rules and states of polarization of the fourteen possible frequencies are derived. The observed bands are assigned to the calculated vibrations and the force constants are evaluated. It is shown that the oscillations involving the C–C valence force are weakly active in contradiction to the theory. In the last section, this activity has been ascribed to the presence of CH_3-groups at the ends of the polythene chain.

Modes of Disintegration of Ionium: an Investigation using β-sensitive Emulsions, by C. JOAN D. JARVIS and M. A. S. ROSS.

ABSTRACT. Kodak NT 4 emulsion has been impregnated in a neutral solution of ammonium ionium oxalate. Observations of α-particle and internal conversion electron tracks are interpreted in terms of excitation levels of the radium nucleus at about 70 kev. (22% intensity) and at about 200 and 270 kev. (combined intensity about 2%). A disintegration scheme is proposed which allocates the known γ-ray of 68 ± 1 kev. to the transition between the two highest levels. The results are reasonably consistent with recent α-ray magnetic analyses and γ-ray absorption measurements.

The Angular Distribution of 3 MeV. Neutrons scattered by Protons and by Deuterons, by W. F. CAPLEHORN and G. P. RUNDLE.

ABSTRACT. The angular distribution of scattering of 3 MeV. neutrons by protons and by deuterons has been investigated with a Wilson cloud chamber. The results reveal an almost isotropic distribution for neutron–proton scattering and a strongly anisotropic distribution in the case of neutron–deuteron scattering. The experimental results are compared with the theoretical curves of Buckingham and Massey for the scattering of 2·5 MeV. neutrons by deuterons.

Flow of Liquid Helium through Fine Channels, by G. K. WHITE.

ABSTRACT. The use of etched copper membranes for studying the flow properties of liquid helium is described. In liquid helium II the velocity of flow is found to be proportional to the *n*th power of the pressure difference, where the exponent *n* has the value 0·22 from 1·15° K. to 1·9° K. and increases slightly near the λ-point. Some determinations of the viscosity of liquid helium I give results in general agreement with earlier work, no indication of non-classical or superfluid flow being apparent either immediately above the λ-point or at higher temperatures.

Pressure Gradients in Superflow, by R. BOWERS and G. K. WHITE.

ABSTRACT. Earlier work on the pressure gradients existing in a narrow channel through which liquid helium II is flowing has been extended by observations on the flow of liquid helium II through porous metallic membranes. Using two membranes of only slightly different flow resistance in series, it has been found that almost all the pressure drop associated with the flow occurs across the membrane of higher resistance, the flow through the membrane of lower resistance taking place under almost zero pressure gradient. The phenomenon is compared with a similar effect in film flow and is interpreted in terms of a critical flow rate in liquid helium II.

Excitation Temperatures of Hydrogen Arcs, by H. EDELS and J. D. CRAGGS.

ABSTRACT. Measurements of the relative intensities of the first three Balmer series lines (H_α, H_β and H_γ) excited in low current (up to 10 amp.) D.C. arcs in hydrogen are described. The pressure range was from 1–2 atmospheres. The results after correction for self-absorption show that the populations of the appropriate excited states are not those given by a Boltzmann distribution, that is, the data obtained cannot be described in terms of a single excitation temperature. Possible causes of this effect are briefly considered.

Balmer Line Widths in Hydrogen Arcs, by H. EDELS and J. D. CRAGGS.

ABSTRACT. The paper describes measurements of the widths of the Balmer lines H_α, H_β and H_γ in D.C. hydrogen arcs (1–2 atmospheres pressure), for currents up to 10 amp. The results are used to deduce ion concentrations, and their bearing on the physical properties of the discharges is discussed.

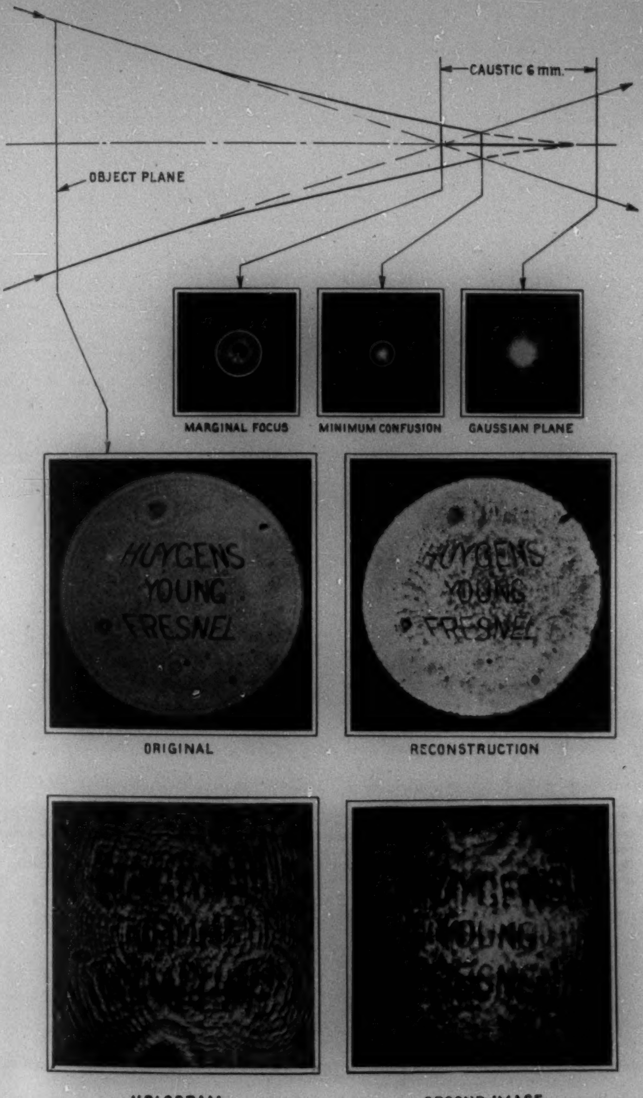

CAUSTIC 6 mm.
OBJECT PLANE
MARGINAL FOCUS
MINIMUM CONFUSION
GAUSSIAN PLANE
HUYGENS
YOUNG
FRESNEL
ORIGINAL
RECONSTRUCTION
HOLOGRAM
SECOND IMAGE

Figure 4. Reconstruction from a hologram taken with the projection method, in the presence of strong spherical aberration. The photographs show three cross sections of the illuminating beam: the original object, the hologram, the reconstructed image and the 'conjugate object'. It is seen that details of the object which are very much finer than the diameter of minimum confusion are reconstructed.

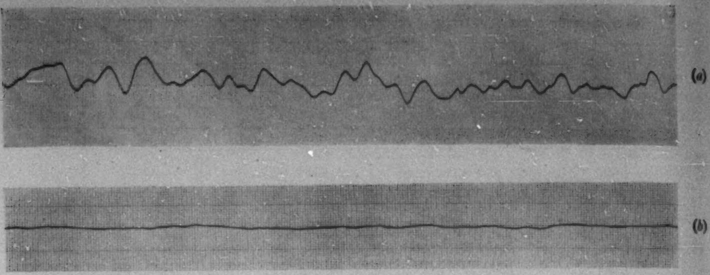


Figure 5. Photographic record, (*a*) showing effect of air currents in disturbing the autocollimating system when there is an unshielded gap of 5 mm. between the autocollimating lens L_2 and the mirror M_2. This record should be compared with (*b*), which was taken under the same conditions but with the gap shielded by a cylindrical sheath extending from lens to mirror. The sensitivity was about $1{\cdot}5 \times 10^{-9}$ radian/scale division in each case, and the primary mirror was fixed. (*b*) shows the residual disturbances which remain in the lever after all reasonable precautions have been taken short of evacuating the light path. Vertical lines are spaced at 1 second intervals.

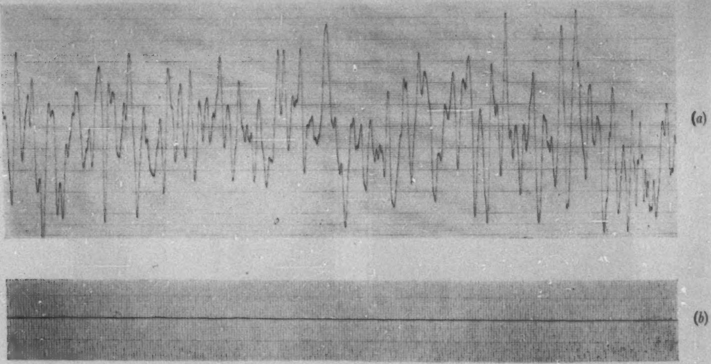


Figure 6. Photographic record, (*a*) showing the Brownian movement of a Tinsley 4500 galvanometer, period 2·7 seconds, in a circuit of resistance 144 ohms with a Kipp and Zonen double coil galvanometer, period 1·6 seconds, as secondary. Record speed about 0·4 mm/sec. (*b*) shows the corresponding record with a fixed primary mirror at the same sensitivity, about $1{\cdot}5 \times 10^{-8}$ radian/scale division. Vertical lines are spaced at 1 second intervals.

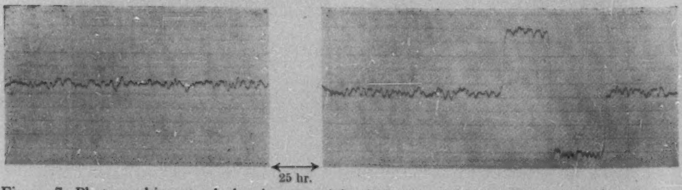


Figure 7. Photographic record showing overnight drift of amplifier and primary galvanometer. Left-hand trace, zero of Tinsley type 4500 galvanometer at 11.00 hr. 8/8/49. Record then stopped and restarted at 12.00 hr. 9/8/49, showing that drift in meantime was small. Calibration current $\pm 1{\cdot}31 \times 10^{-10}$ amp. Primary circuit resistance 150 ohms.

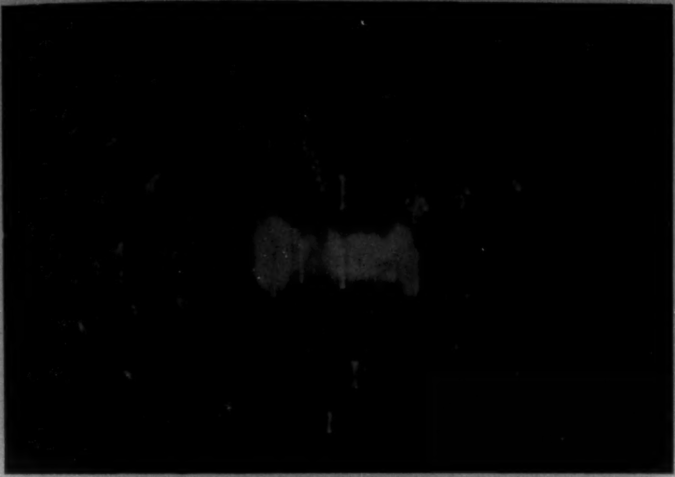


Figure 2. Reed of width 0·14 cm. vibrating in smoke-laden air. Frequency, 3·3/sec. Double amplitude 0·035 cm.

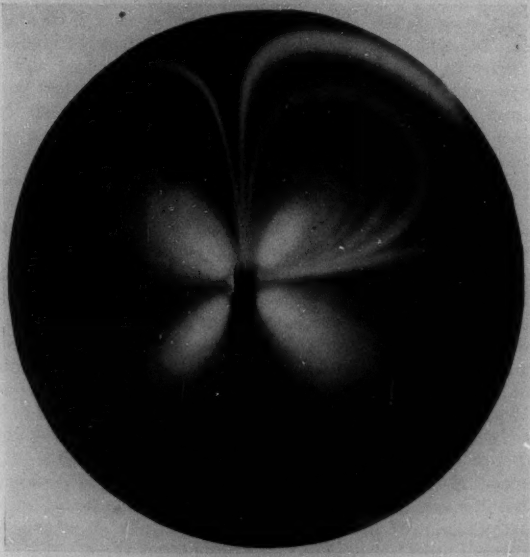


Figure 4. Circulation around a vibrating reed. The width of the reed (shown by the vertical dark shadow) is 0·16 cm. and the frequency 50 per sec. The double amplitude is 0·11 cm.